Présentation du projet

Digital Workflow from Text Retrieval to Text Alignment on a Medieval French Text

Lucence Ing

Centre Jean Mabillon, École des chartes – PSL

28 mars 2022

1. Introduction

Objectifs

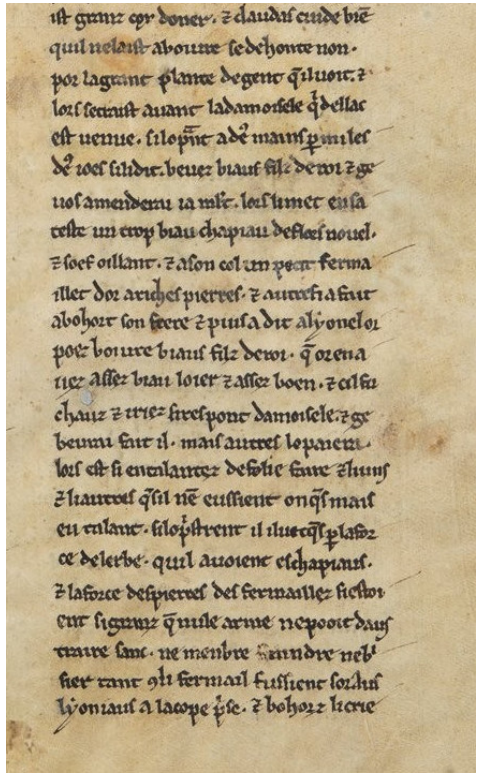
Objectifs de l'atelier

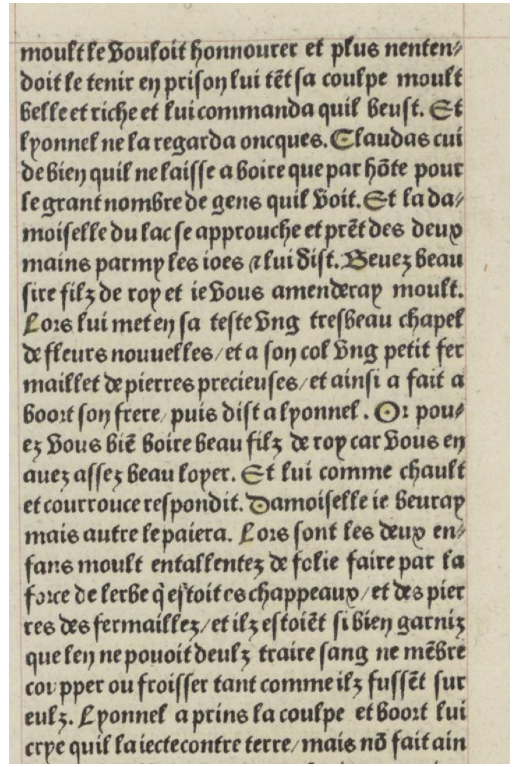
- faire **découvrir** un certain nombre de **méthodes et d'outils** DH/DL
- faire **pratiquer** ces méthodes et outils
- comprendre comment les outils sont **implémentés au sein d'un projet de recherche**

Objectifs du projet

- **aligner plusieurs versions** d'un même texte pour les comparer
- réaliser cet alignement de manière **automatique**
- plusieurs étapes sont nécessaires

De l'image aux données textuelles





Ao	Ez
por	pour
la	le
grant	grant
planté	nombre
de	de
gent	genz

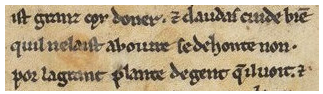
Trois étapes (1)

1. récupération des données textuelles : ***Handwritten Text Recognition*** (Reconnaissance de l'écriture manuscrite)
2. enrichissement des données : **annotation linguistique** des données (**Traitement automatique des langues** (TAL ou *NLP*, *Natural Language Processing*))
3. alignement des textes annotés : **collation automatique**

Planning des étapes :

lundi	présentation du projet et ajustement selon les envies
mardi	HTR
merc	annotation linguistique
jeudi	collation
vendr	finition des données et préparation de la présentation

Trois étapes (2)

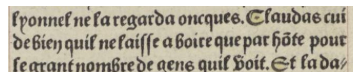


HTR

por la grant planté

annotation

<w lemma="por2">por</w>



HTR

pour le grant nombre

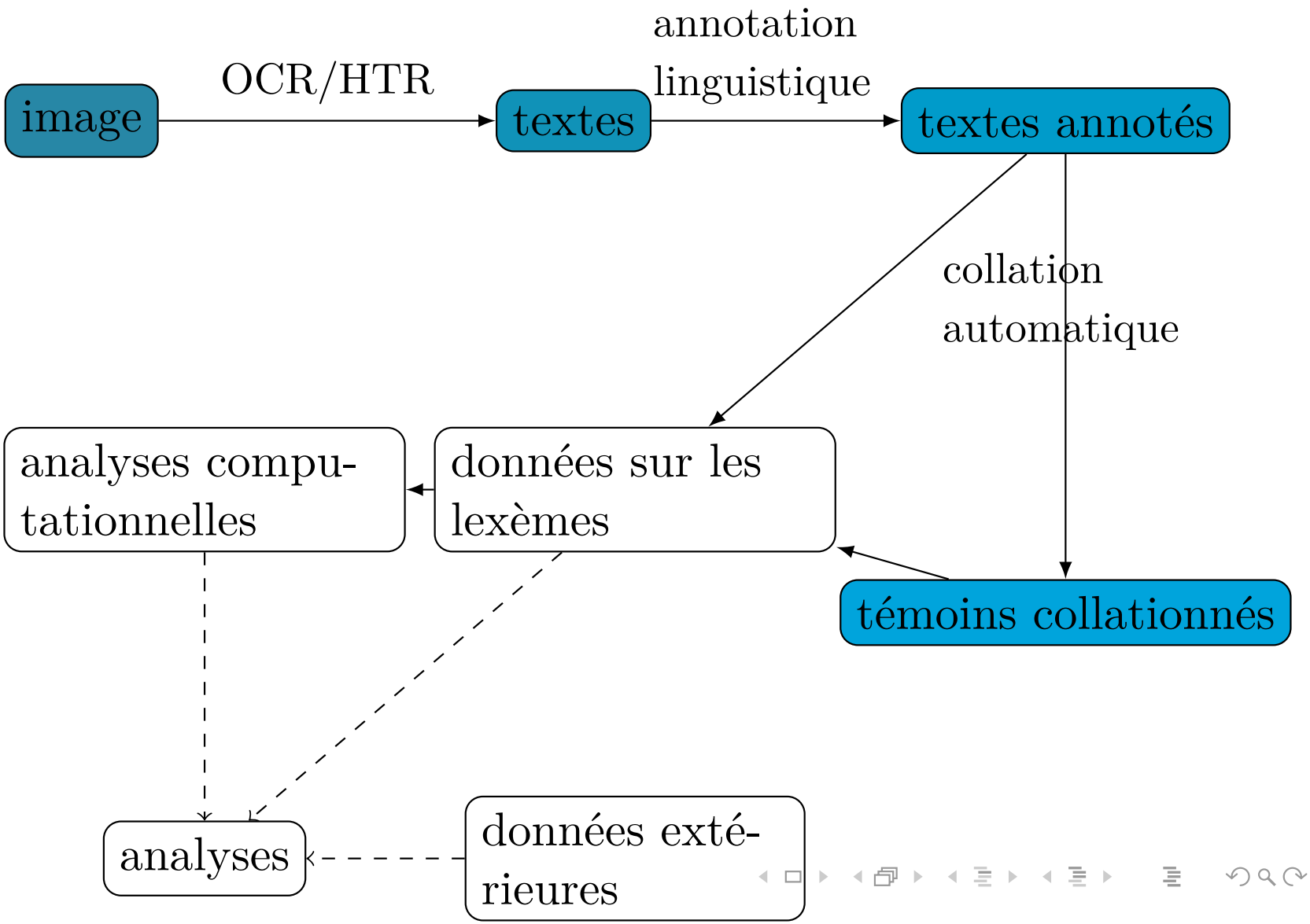
annotation

<w lemma="por2">pour</w>

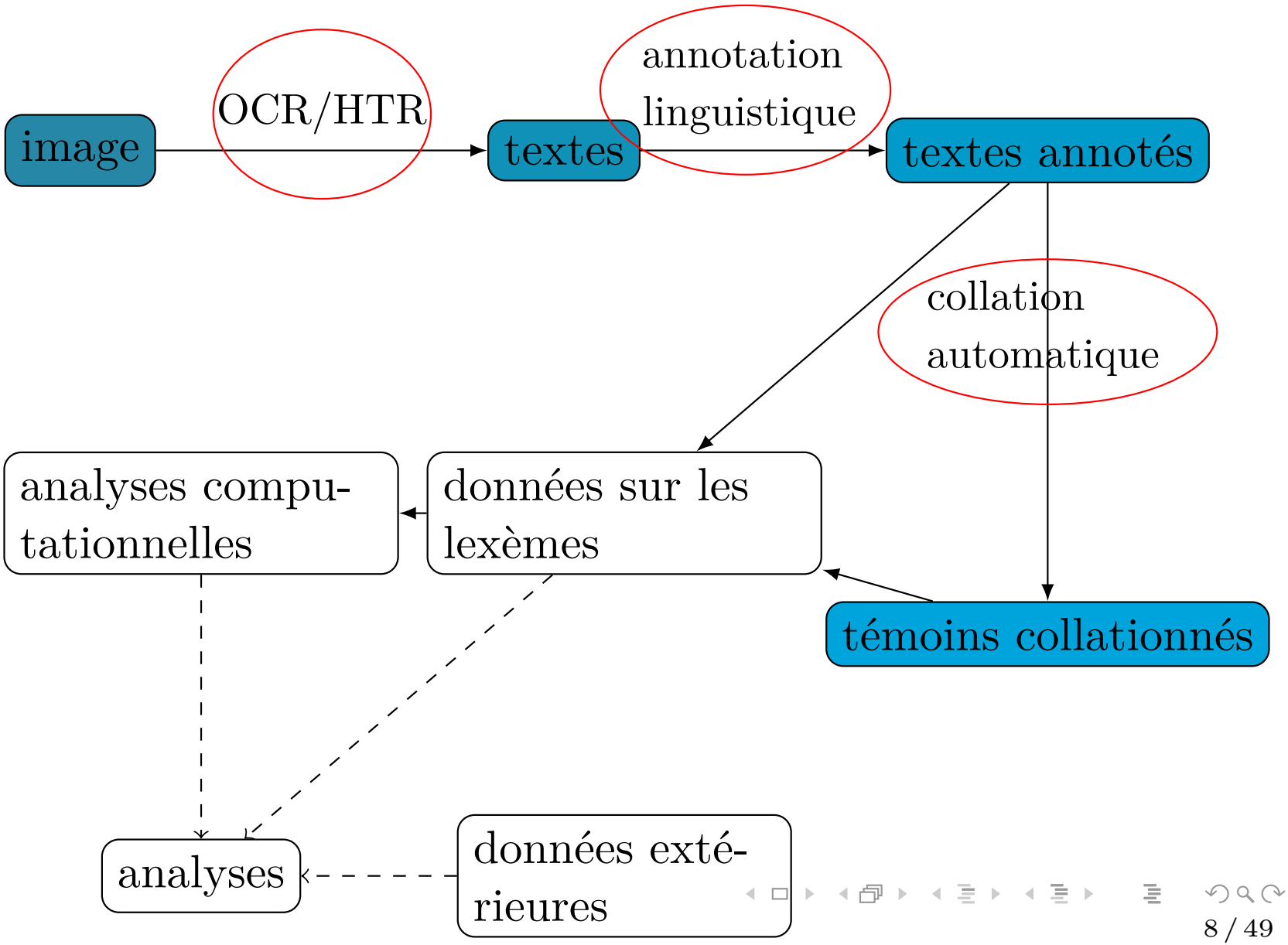
collation

Ao	Ez
por	pour
la	le
grant	grant
planté	nombre
de	de
gent	genz

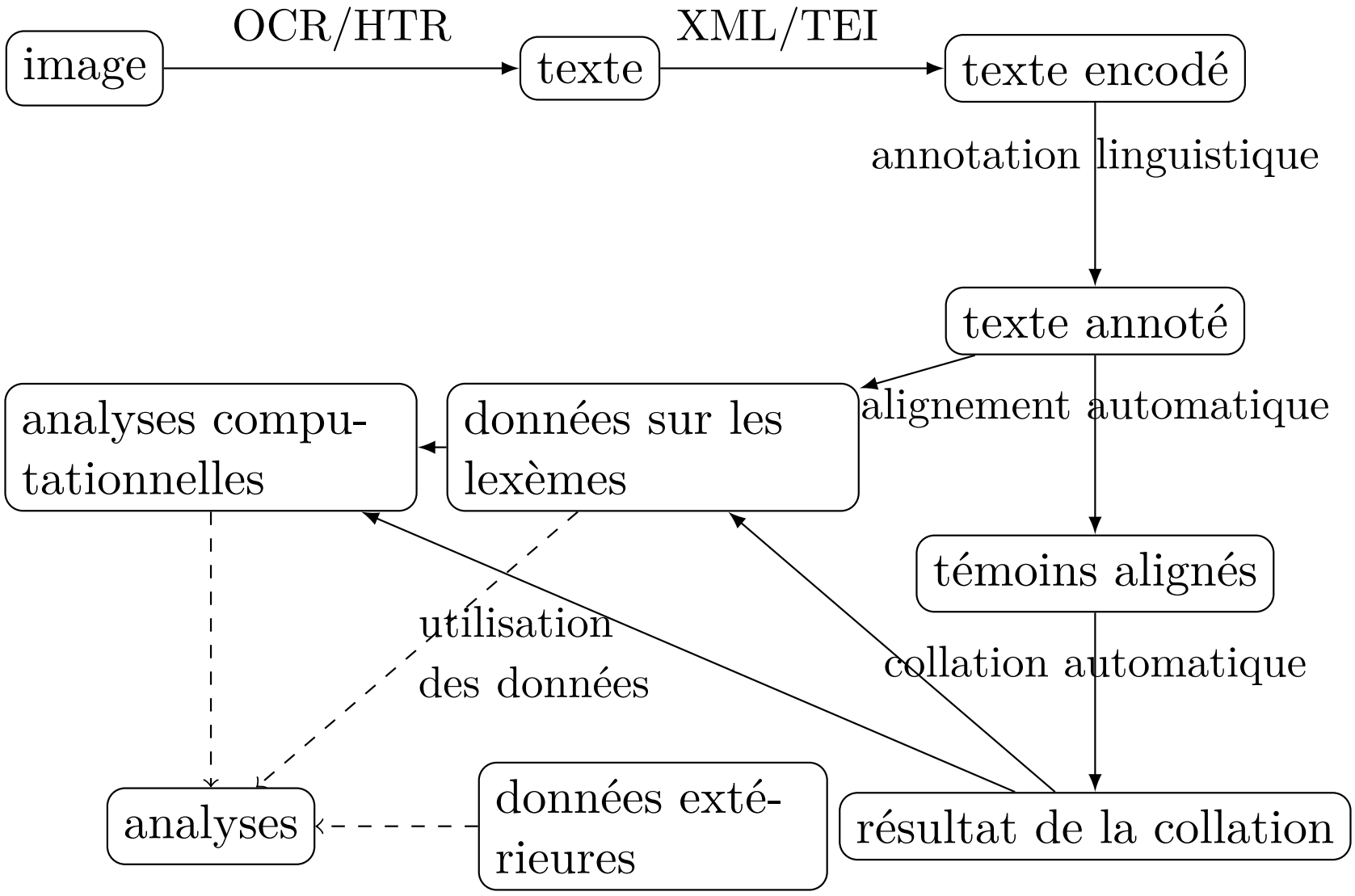
Chaîne des trois étapes



Chaîne des trois étapes



Chaîne de traitements complète



Français médiéval

Une langue non standardisée

chaque mot peut être écrit de **plusieurs manières** (plusieurs graphies)

exemples des graphies du mot **enfant** : *enfans*, *anfanz*, *enfant*, *anfant*, *anfes*, *enfes*, *anffans*, *enffant*, *amfant*, *enfanz*

Français médiéval

Une langue non standardisée

chaque mot peut être écrit de **plusieurs manières** (plusieurs graphies)

exemples des graphies du mot **enfant** : *enfans*, *anfanz*, *enfant*, *anfant*, *anfes*, *enfes*, *anffans*, *enffant*, *amfant*, *enfanz*

Variation des textes

- **le processus de la copie** entraîne des modifications dans le texte (des erreurs par exemple)
- les textes peuvent être **modifiés volontairement** par les scribes

  Ao De ceste avanture s'esbaudirent mout. Et si les avoit la perriere si estoutoiez et les murs peçoiez et estonez. Et Claudas apela un jor Banyn.
  Ez De ceste avanture moult s'esbaudirent. Et Claudas appella ung jour Banyn.

Quatre versions de l'*incipit*

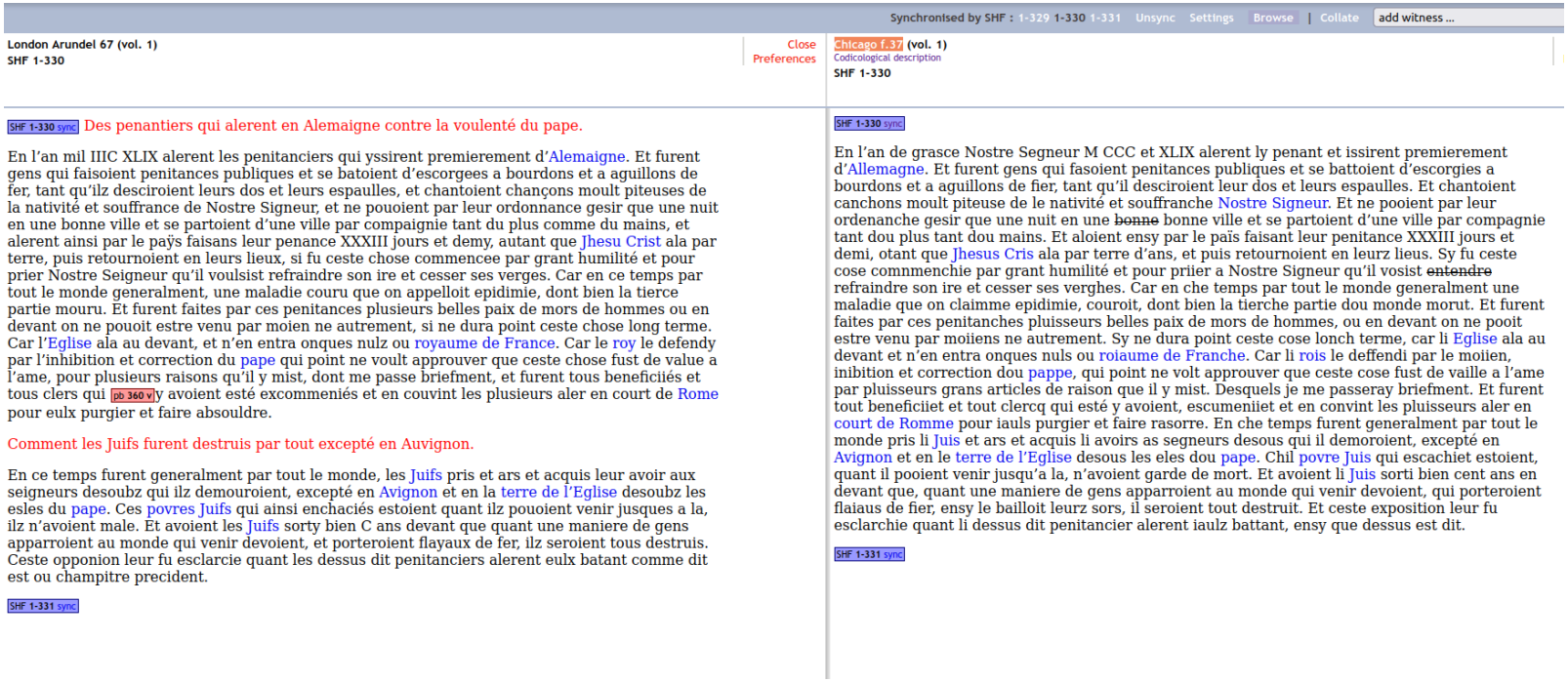
K En la marche de Gaule et de la Petite Bretaigne avoit deus rois **anciennement** qui estoient frere germain et avoient deus **serors** germaines a fames.

Ez En la marche de Gaule et de la Petite Bretaigne avoit **anciennement** deux roys freres germains, si avoient a femmes deux **seurs** germaines.

Ao En la marche de Gaule et de la Petite Bretaigne avoit .ii. rois **anchienement** qui estoient freire germain et avoient a femmes .ii. **serours** germaines.

R En la marche de Gaule et de la Petite Bretaingne avoit .ii. rois **encienement** qui estoient frere germain.

Exploitations : édition



Voir le Projet Online Froissart :
`https://www.dhi.ac.uk/onlinefroissart/apparatus.jsp?type=summary`

Exploitation : étude linguistique

bëer

Onques par amors n'avoit amé c'une foiee et qant l'an li demandoit por quoi il avoit aimors laissiees, si disoit que por ce qu'il baoit a vivre longuement.

Oncques n'avoit aymé fors une fois et quant on lui demandoit pourquoy il avoit amours laissees, il disoit qu'il desiroit vivre longuement.

... ele ot fait grant partie de ce que ele baoit affaire, si fu mout liee et petit prisa lo cop...

... elle eust fait grant partie de ce qu'elle vouloit faire, elle fut moult joyeuse et peu prisa le coup...

Le texte du *Lancelot*

Le texte

- anonyme, premier tiers du XIII^{e} siècle
- **partie centrale** du *Lancelot-Graal*
- un texte **très diffusé** jusqu'au premier tiers du XVI^{e} siècle

Les **versions**

- version longue et version courte
- version de Londres et version de Paris à l'intérieur de la version longue
- contaminations diverses

La tradition du texte

Quant à vouloir fixer strictement la place des manuscrits les uns par rapports aux autres et ***dessiner un de ces beaux arbres généalogiques*** *dont s'ornent les éditions critiques, il* ***n'y faut pas songer.***
(A. Micha, « La Traduction manuscrite du *Lancelot* en prose », 1964)

Nos témoins

- *Lancelot* traditionnellement divisé en trois parties : le **Galehaut**, la Charette et l'Agravain
- les versions commencent à diverger lors l'épisode du deuxième voyage en Sorelois (fin du Galehaut)
- notre étude porte sur la **partie similaire** du *Lancelot*

Extrait retenu

Chapitre 011 du *Lancelot* en prose (d'après le découpage de l'édition de E. Kennedy)

Témoins retenus :

- **Rennes BM, ms 255** (premier tiers XIII^{e}, *scripta* centrale). `https://bvmm.irht.cnrs.fr/consult/consult.php?mode=vignettes&reproductionId=925&VUE_ID=-1&`. Passage du f. 150vb au f. 152vc.
- **BnF, fr. 16999** (milieu XIV^{e}, Paris) `https://archivesetmanuscrits.bnf.fr/ark:/12148/cc13609v`. Passage du f. 17rc au f. 19vc.
- **BnF, fr. 113-4** (dernier quart XV^{e}, Poitiers) `https://gallica.bnf.fr/ark:/12148/btv1b60000903`. Passage du f. 168ra au f. 170va sur le BnF, fr. 113.

Les défis

Un objectif de recherche

- textes annotés : nombreuses **exploitations** possibles
- avoir les textes alignés : repérer automatique-ment les **lieux variants**

Un défi computationnel

- HTR : écriture **manuscrite** ; trois siècles différents
- annotation linguistique : multiplicité des **graphies**
- alignement : des **passages variants**

➡ **hétérogénéité des données**

L'apprentissage profond (1)

Deep-learning

Le **principe** est simple : **apprendre à un modèle à traiter des données**.

Fonctionnement sur des réseaux de neurones (moins simple !).

Les composantes

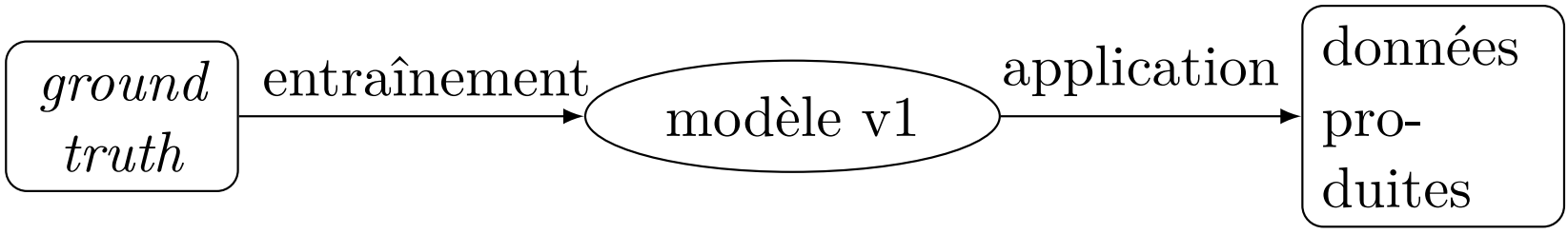
1. une **tâche** (par ex., produire un texte à partir de l'image de ce texte)
2. des **données annotées** (le résultat de ce que doit produire le modèle) : le *ground truth* ou la vérité-terrain
3. un **modèle**

L'apprentissage profond (2)

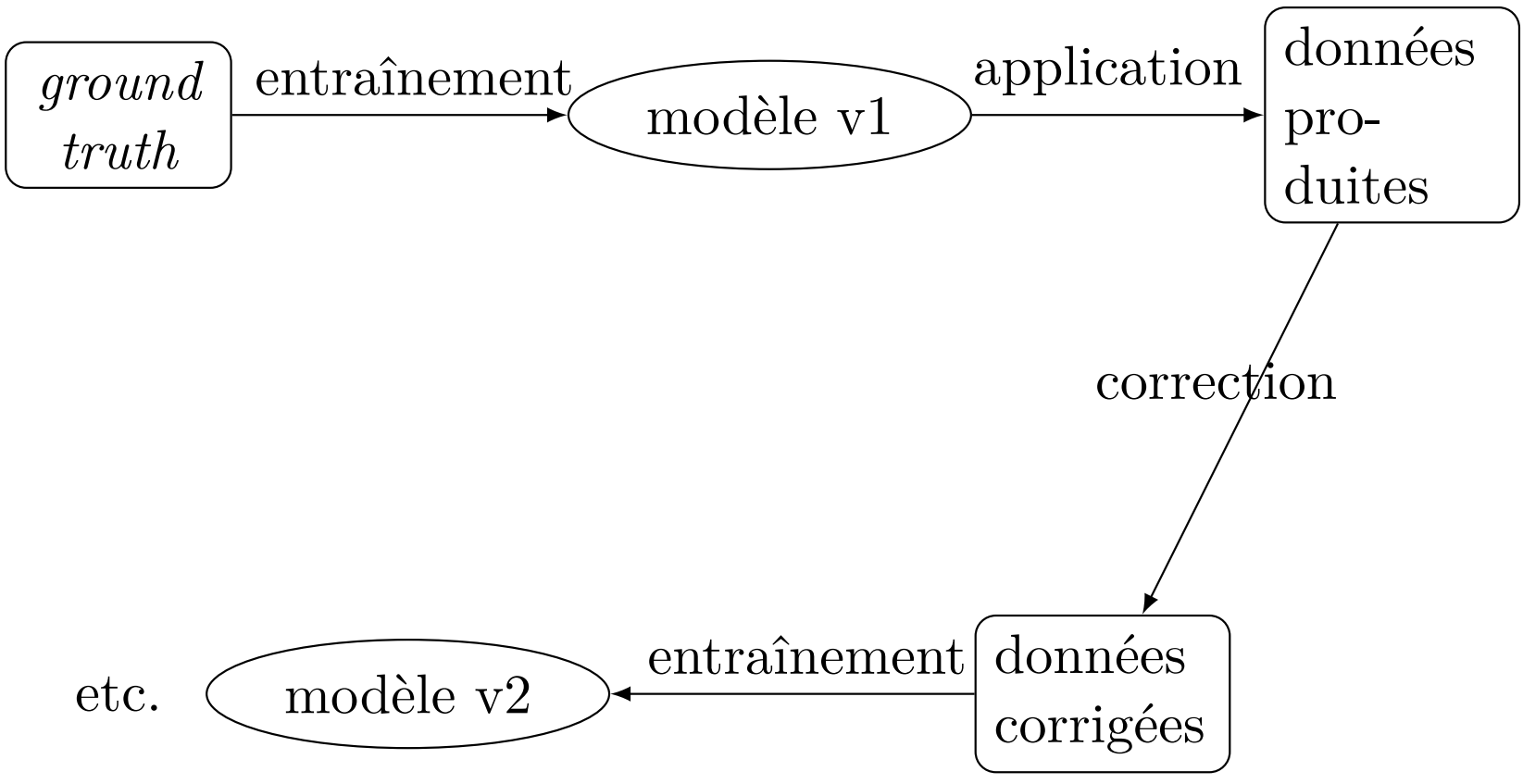
Fonctionnement

- un **modèle**...
- ...qu'on **entraîne** sur des données annotées...
- ...qu'on **applique** sur d'autres données,...
- ...résultats que l'on **corrige**...
- ...pour **entraîner à nouveau** le modèle...
- ...qu'on **applique à nouveau** sur d'autres données...
- ...dans une sorte de boucle vertueuse d'**amélioration du modèle**

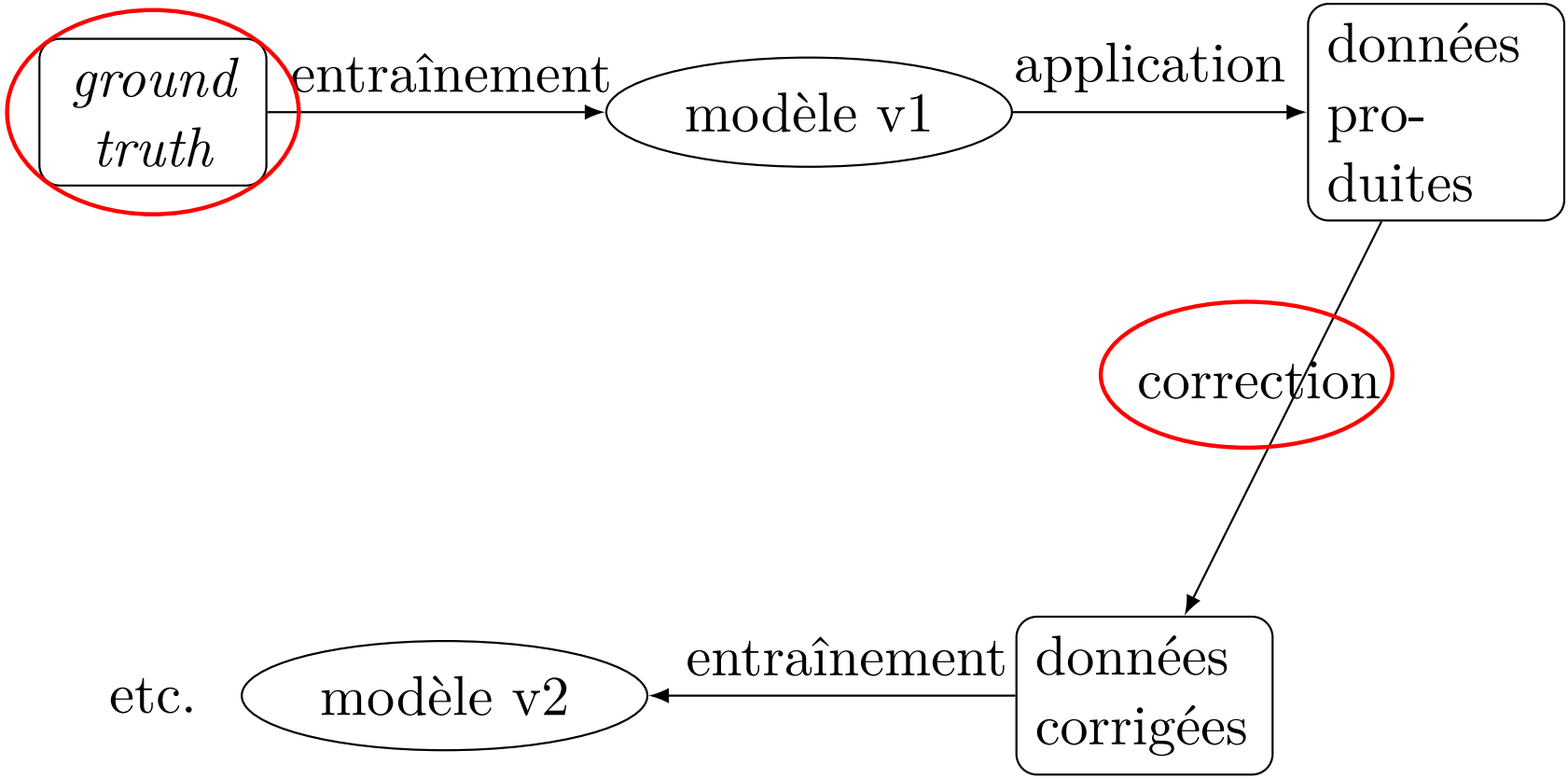
L'apprentissage profond (3)



L'apprentissage profond (3)



L'apprentissage profond (3)

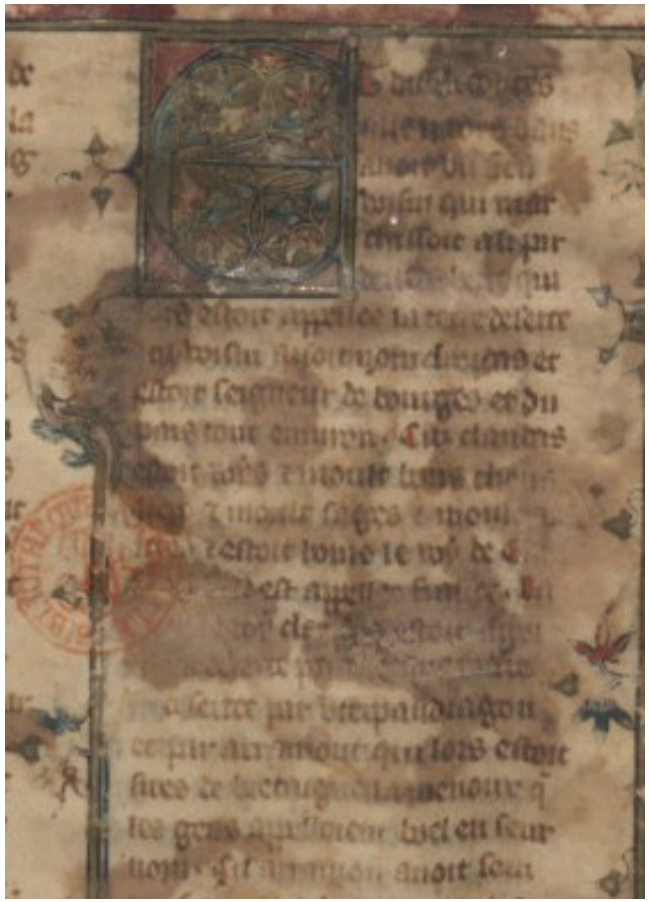


2. Récupération des données

HTR

Handwritten Text Recognition

Technologie plutôt récente ; basée sur le principe de l'OCR (*Optical Character Recognition*)

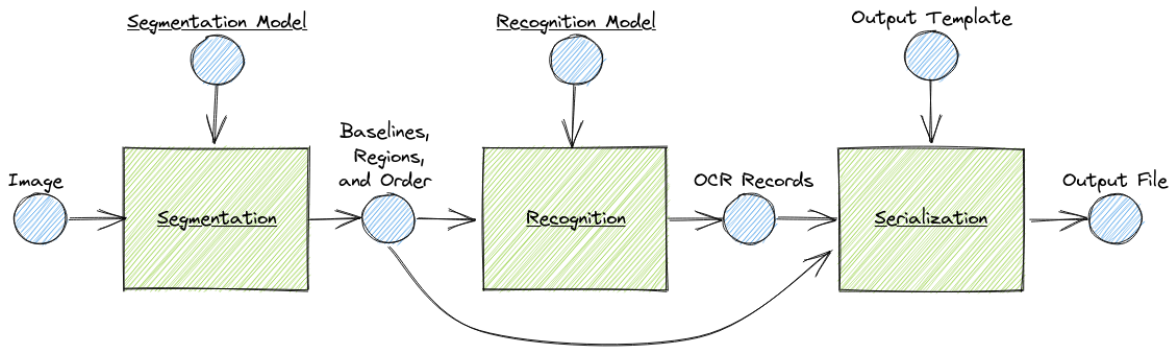


Difficultés

- **irrégularité** de l'écriture manuscrite
- **état matériel** des documents (surtout pour les documents anciens)

kraken

Système de reconnaissance des caractères adapté pour la reconnaissance de l'écriture manuscrite et pour la reconnaissance des écritures non-latines (caractères non-latins, sens de lecture différent).



Source image : site de `kraken`

Site : `https://kraken.re/master/index.html`
Voir le projet HTR-United : `https://htr-united.github.io/`

escriptorium (1)

Une application **open source** qui permet de réaliser l'HTR dans une **interface conviviale**.
Développée par le projet Scripta (PSL).

- découpage des **zones de texte**
- découpage des **zones de ligne**
- **reconnaissance** des caractères

Documentation escriptorium : `https://lectaurep.hypotheses.org/documentation/escriptorium-tutorial-en`

escriptorium (2)

Étapes permises

- création de **vérité terrain**
- **entraînement** de modèles (**segmentation** et **reconnaissance**)
- **correction** des données

Fonctionnalités

- import et export des **données**
- import et export des **modèles**
- **partage** des données et des modèles
- multiplicité des formats supportés

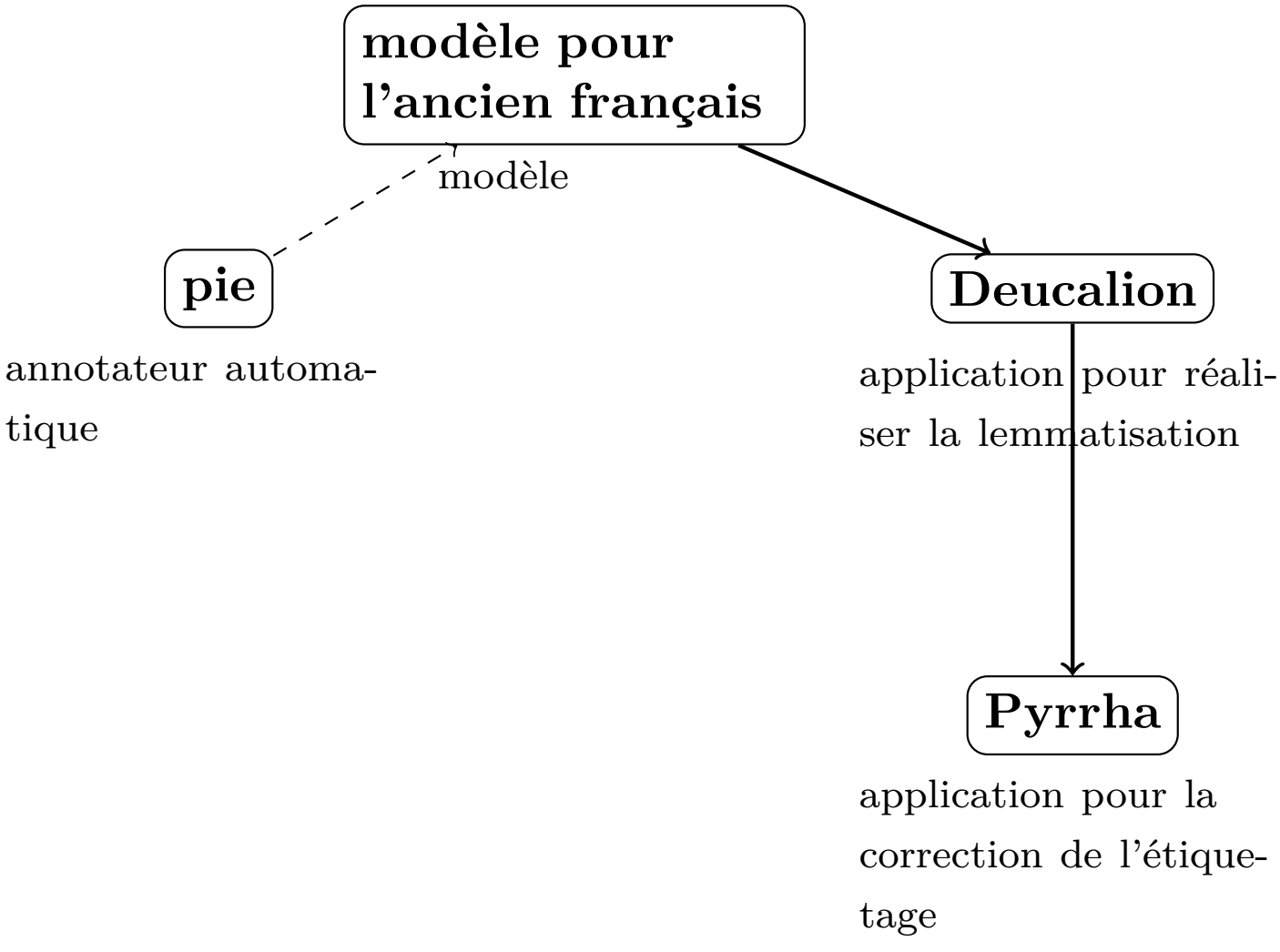
3. Annotation des données

Annotation linguistique : outils

Annotation et correction

- annotateur automatique : **pie** (E. Manjavacas, T. Clérice, M. Kestemont)
  https://doi.org/10.5281/zenodo.4572585
- interface Deucalion (T. Clérice)
  https://dh.chartes.psl.eu/deucalion/
- **modèle pour l'ancien français** entraîné à l'École nationale des chartes (J.-B. Camps et al.)
- correction avec l'interface Pyrrha (T. Clérice, J. Pilla et al.)
  https://doi.org/10.5281/zenodo.5144781

Schéma des outils d'annotation linguistique



Annotation linguistique : principes

Étiquettes retenues pour le projet : **lemmes** et **partie du discours** (POS, *part-of-speech*)

Modèle pour l'ancien français

lemme dictionnaire Tobler-Lommatzsch
`https://www.ling.uni-stuttgart.de/institut/ilr/toblerlommatzsch/work/workfr.htm`

POS en suivant la documentation Cattex
`http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_manuel_2.0.pdf`

Ao	ses	hiaumes	fu	toz	fanduz
Ez	son	heaume	fut	tout	fendu
lemma	son4	heaume	estre1	tot	fendre
pos	DETpos	NOMcom	VERcjg	ADVgen	VERppe

Annotation dans le fichier XML

```
<w xml:id="Ao_w_0135673" lemma="il" pos="PROper">il</w>
<w xml:id="Ao_w_0135674" lemma="ne1" pos="ADVneg">n</w>
<w xml:id="Ao_w_0135675" lemma="avoir" pos="VERcjg">eüssient</w>
<w xml:id="Ao_w_0135676" lemma="mie" pos="ADVneg">mie</w>
<w xml:id="Ao_w_0135677" lemma="vëoir" pos="VERppe">veü</w>
<w xml:id="Ao_w_0135678" lemma="le" pos="DETdef">lo</w>
<w xml:id="Ao_w_0135679" lemma="lïon" pos="NOMcom">lion</w>
<w xml:id="Ao_w_0135680" lemma="en1" pos="PRE">en</w>
<w xml:id="Ao_w_0135681" lemma="le" pos="DETdef">l</w>
<w xml:id="Ao_w_0135682" lemma="aigue" pos="NOMcom">eive</w>
<w xml:id="Ao_w_0135683" lemma="mais1" pos="CONcoo">mais</w>
<w xml:id="Ao_w_0135684" lemma="làsus" pos="ADVgen">laïssus</w>
<w xml:id="Ao_w_0135685" lemma="en1+le" pos="PRE.DETdef">el</w>
<w xml:id="Ao_w_0135686" lemma="ciel" pos="NOMcom">ciel</w>
<w xml:id="Ao_w_0135687" lemma="car" pos="CONcoo">Car</w>
<w xml:id="Ao_w_0135688" lemma="le" pos="DETdef">li</w>
<w xml:id="Ao_w_0135689" lemma="ciel" pos="NOMcom">ciaus</w>
<w xml:id="Ao_w_0135690" lemma="estre1" pos="VERcjg">est</w>
<w xml:id="Ao_w_0135691" lemma="siecle" pos="NOMcom">siegles</w>
<w xml:id="Ao_w_0135692" lemma="pardurable" pos="ADJqua">pardurables</w>
```

La force des lemmes

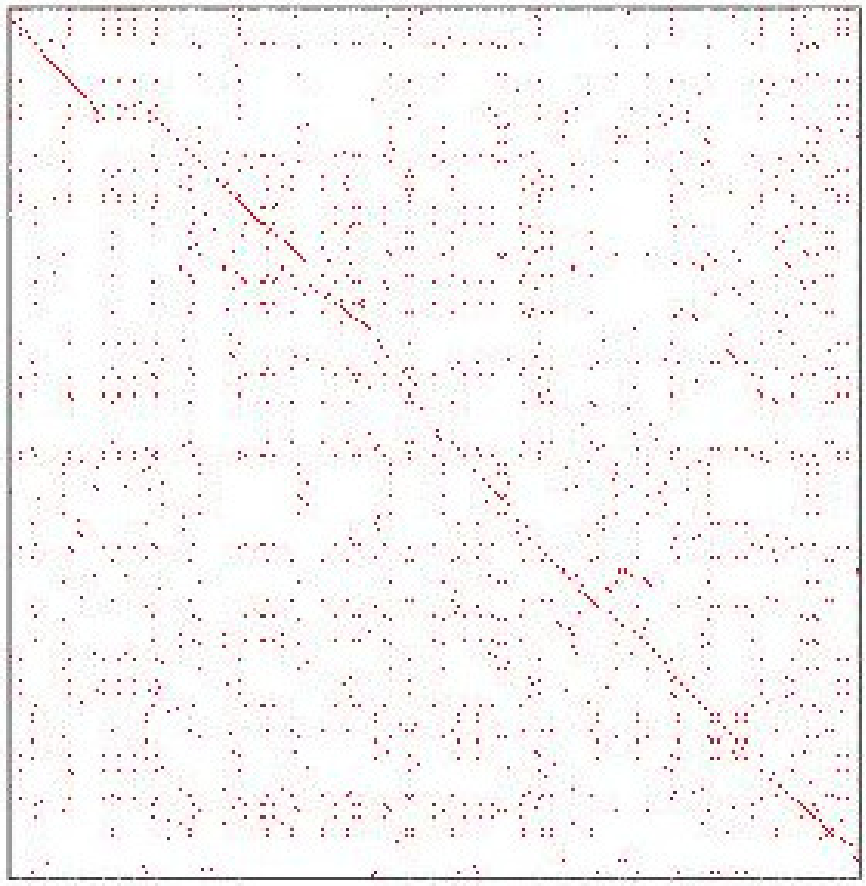
Pourquoi lemmatiser ?
→ obtenir des **formes stables** : résoudre le problème de variation de graphies

10 formes pour **enfant** dans notre corpus : *enfans*, *anfanz*, *enfant*, *anfant*, *anfes*, *enfes*, *anffans*, *enffant*, *amfant*, *enfanz*

- faire des calculs de **fréquence**
- pouvoir rechercher des **usages** de manière simple
- permettre un **alignement** sur des **lemmes identiques**

4. Alignement des données

Aligner les témoins



Matrice représentant l'alignement de deux témoins sur le chapitre 5. Chaque point rouge représente un valeur de distance à 0 (tokens identiques)

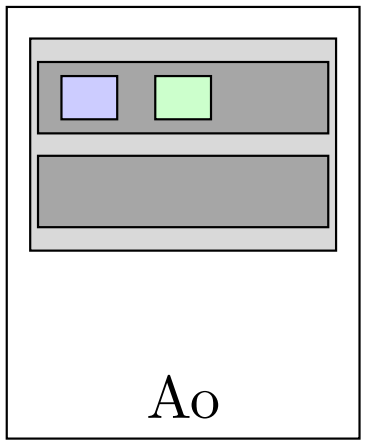
Matrice de distance

	cant1	le	chevalier	desireter	öir	le	novele
le	5	0	7	8	3	0	4
conte1	3	5	7	7	6	5	5
dire	5	3	8	5	2	3	5
que4	5	3	8	8	4	3	5
cant1	0	5	7	8	5	5	6
le	5	0	7	8	3	0	4
chevalier	7	7	0	7	7	7	6
desireter	8	8	7	0	7	8	7
öir	5	3	7	7	0	3	6
le	5	0	7	8	3	0	4
novele	6	4	6	7	6	4	0
Montlair	6	7	7	8	6	7	6

Exemple de la matrice de distance, sur le début du chapitre 5

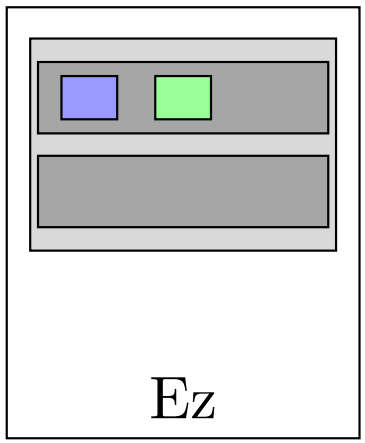
J.-B. Camps, E. Spadini, L. Ing, FALCON, “Collating Medieval Vernacular Texts. Aligning Witnesses, Classifying Variants”, `https://hal.archives-ouvertes.fr/hal-02268348`

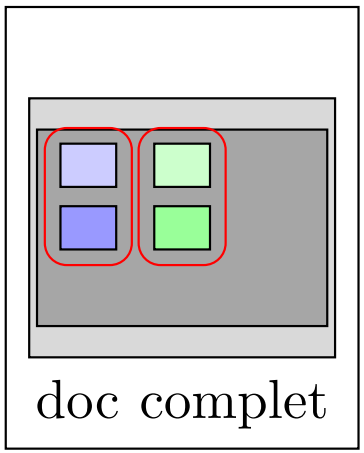
Collation



la collation automatique produit un alignement mot à mot

il est fait sur les lemmes grâce à Collatex





Fonctionnement de la collation (1)

La **collation**

- étape dans l'**édition d'un texte** : repérer les différentes variantes des témoins d'un texte afin d'en établir la tradition textuelle
- permet aussi de **comparer les témoins**

Collatex

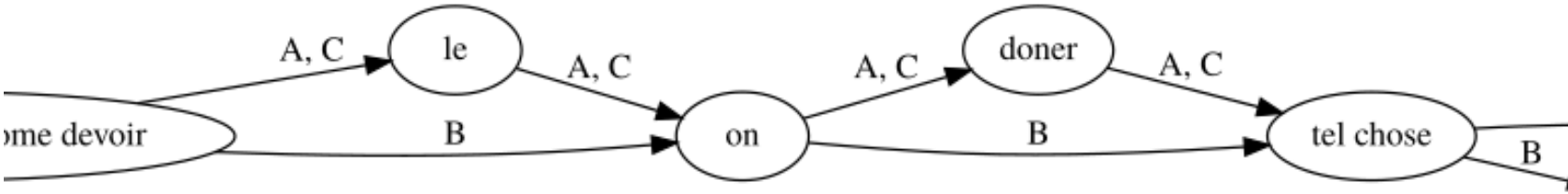
- un outil de **collation automatique**
- permet d'**aligner** les tokens et d'identifier similarités et différences entre les témoins
- l'alignement est **basé sur l'identification de chaînes de caractères similaires** dans deux témoins ou plus
- à cause de la **variance** en français médiéval, la collation est réalisée sur les **lemmes**

Fonctionnement de la collation (2)

Modèle Gothenburg

Modèle défini en 2009, composé de **quatre étapes** :

1. tokénisation
2. alignement (à partir d'un témoin, témoin après témoin)
3. détection des transpositions
4. visualisation



https://collatex.net/
http://interedition.github.io/collatex/pythonport.html

Exemple de collation (vue table)

Ao	Ez
Li	-
contes	-
dit	-
que	-
qant	Quant
li	le
chevaliers	chevalier
deseritez	desherité
oï	ouyt
les	les
noveles	nouvelles
-	de
Monlair	Moncler
lo	-
chastel	-
qui	qui
pris	prins
estoit	estoit
et	et
il	il
vit	vit
Claudas	Claudas

Exemple de collation (vue XML)

```
<app type="graph">
<rdg wit="#Ao">
<w lemma="öir" xml:id="Ao_w_0009258" pos="VERcjg">oï</w></rdg>
<rdg wit="#Ez">
<w lemma="öir" xml:id="Ez_w_0013454" pos="VERcjg">ouyt</w></rdg>
</app>
<app type="absVar">
<rdg wit="#Ao">
<w lemma="le" xml:id="Ao_w_0009259" pos="DETdef">les</w></rdg>
<rdg wit="#Ez">
<w lemma="le" xml:id="Ez_w_0013455" pos="DETdef">les</w></rdg>
</app>
<app type="graph">
<rdg wit="#Ao">
<w lemma="novele" xml:id="Ao_w_0009260" pos="NOMcom">noveles</w></rdg>
<rdg wit="#Ez">
<w lemma="novele" xml:id="Ez_w_0013456" pos="NOMcom">nouvelles</w></rdg>
</app>
<app type="leconIsolee" corresp="#Ez">
<rdg wit="#Ez">
<w lemma="de" xml:id="Ez_w_0013457" pos="PRE">de</w></rdg>
</app>
<app type="graph">
<rdg wit="#Ao">
<w lemma="Montlair" xml:id="Ao_w_0009261" pos="NOMpro">Monlair</w></rdg>
<rdg wit="#Ez">
<w lemma="Montlair" xml:id="Ez_w_0013458" pos="NOMpro">moncler</w></rdg>
</app>
```

Les erreurs de collation

chap	n°p	id. l	préc. Ao	préc. Ez
divColl62	122	Ao_w_0232114-Ez_w_0172535	que vos an vandroiz par moi sanz autre bessoigne	damoiselle que vous en revendrez par moy sans autre affaire
divColl63	41	Ao_w_0242661-Ez_w_0182317	por vostre bessoigne Mais ge ne la cuidai pas avoir	pour vostre besongne et je ne la cuidoie pas cy

l. Ao	l. Ez	suiv. Ao	suiv. Ez
anprandre	entreprendre	Oïl fait il se ce n estoit afaires dom ge	Oy dit il se se n est chose donc je
amprise	endroit	si a droit com ge ai Mais or sai	avoir entreprinse comme j ay mais ore scay je

5. Autres approches computationnelles

Deux exemples d'approches utilisées

- le **Topic Modeling**, qui permet de déterminer un certain nombre de **thèmes au sein d'un corpus**, et de relier ces thèmes aux différentes occurrences d'un lexème
- le **Word Embeddings**, ou plongement de mots, qui permet d'obtenir un « espace sémantique » des textes, en se basant sur la représentation de **chaque mot comme un vecteur au sein d'une matrice**

Des approches qui permettent de parler du **sémantisme** des lexèmes, basées sur le principe qu'un mot prend sens en contexte. J. Rupert Firth (1957) : « ***You shall know a word by the company it keeps*** ».

Le Topic Modeling

Interêts

- déterminer un **nombre de thèmes** qui structurent le texte
- permet de dessiner des **espaces sémantiques**, sans avoir à assigner manuellement ces espaces (P. Schöch, 2012)
- assigner un **thème à chaque occurrence** : est-ce que certains thèmes sont davantage concernés que d'autres par le processus d'obsolecence ?

→ tout document textuel est constitué de thèmes qui le structurent, et chacun de ces thèmes se caractérise par les mots qui permettent son expression

Word Embeddings

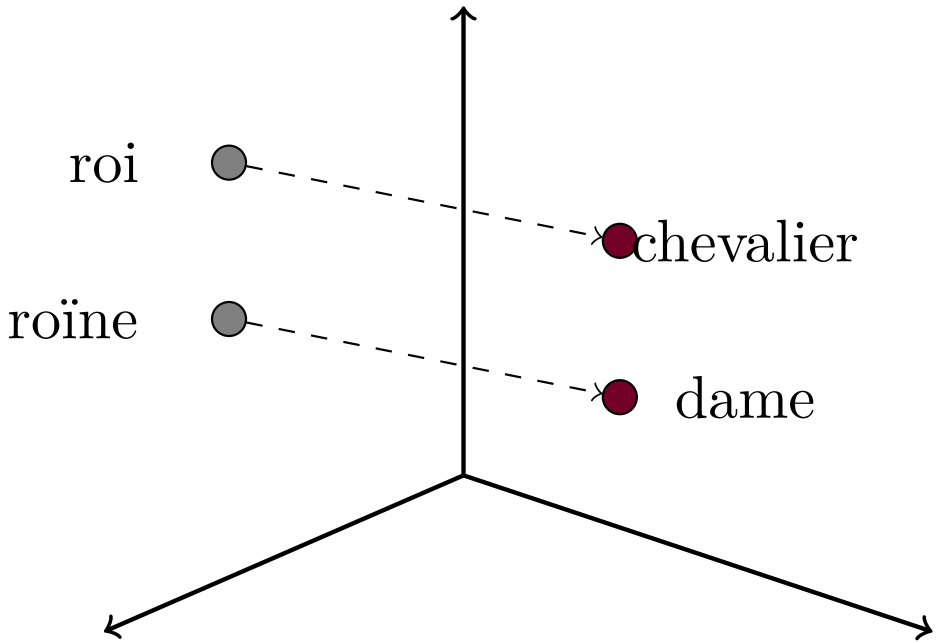
Principe de fonctionnement

- le **texte** est représenté sous une forme de **matrice**
- la **représentation d'un mot** se fait à l'intérieur d'un **vecteur** : le mot est un vecteur de chiffres
- ce vecteur est composé de valeurs établies **en fonction de la co-occurrence** du lexème avec les autres termes

chevalier	-1.5488147	-1.117865	-0.13586469	-0.9344863	1.6124617
roi2	-3.944645	-1.7473782	0.25746804	-2.7548037	2.2756793
dame	0.6642276	-0.61475843	0.00583692	1.5918058	2.1516817

Représentations

Les **mots avec des vecteurs similaires sont des mots qui apparaissent dans des contextes similaires** donc qui ont des sens similaires



Les modèles qui représentent bien les textes peuvent résoudre des équations du type : `roi + dame - chevalier ?`
$\rightarrow$ `roïne`

À vous de jouer !

Questions ? Envies particulières ?