



**HAL**  
open science

## Cluster 7 Biblissima+ projets en cours

Jean-Baptiste Camps, Matthias Gille Levenson, Sébastien Hamel, Lucence Ing, Alice Leflaïc, Laurence Mellerin, Luc Massip, Emmanuelle Morlock, Jules Nuguet, Philippe Pons

► **To cite this version:**

Jean-Baptiste Camps, Matthias Gille Levenson, Sébastien Hamel, Lucence Ing, Alice Leflaïc, et al.. Cluster 7 Biblissima+ projets en cours. Journées annuelles Biblissima+ 2024, May 2024, Paris, France. Zenodo, 2024, 10.5281/zenodo.11402844 . hal-04753020

**HAL Id: hal-04753020**

**<https://hal.science/hal-04753020v1>**

Submitted on 6 Dec 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Jean-Baptiste Camps, Matthias Gille Levenson, Sébastien Hamel, Lucence Ing, Alice Leflaëc, Luc Massip, Laurence Mellerin, Emmanuelle Morlock, Jules Nuguet, Philippe Pons

## Premiers chantiers 2024

Le cluster 7 oriente ses travaux sur **l'interopérabilité, l'alignement et l'analyse des textes**. En plus du travail en cours sur DTS, cette année, l'équipe s'est donné comme objectif premier la constitution d'une base lexicale commune dans data.biblissima, dont la structure modulable permet le traitement **de langues et d'époques multiples**. Des **alignements de référentiels** sont en cours.

Un autre chantier concerne **l'alignement automatique de textes**. L'équipe vise à développer un outil permettant l'alignement de témoins de **différentes langues**, voire leur collation, à l'aide de modèles de langue. Elle s'intéresse, dans un premier temps, aux traditions médiévales.

### Alignement multilingue

**Approche originelle** : alignement et collation (*collatex*) automatiques basés sur des *n-grams*, des **chaînes de caractères identiques**

Ao	Li	contes	dit	que	qant	li	chevaliers	deseritez	oï
Ez	-	-	-	-	Quant	le	chevalier	deshérité	ouyt

Exemple simple de collation sous format tabulaire

→ nécessité de lemmatiser les textes, du fait de la forte variation graphique des langues vernaculaires médiévales

### Nouvelle approche

- travail sur des **traditions multilingues** (*De Regimine Principum* en castillan méd. et en latin ; *Lancelot en prose* en français, castillan et italien médiévaux)
- une **approche sémantique**, qui utilise des modèles de langue multilingues (*bertalign* pour l'alignement, *awesome-align* pour la collation [à venir]), après un travail de tokénisation (entraînement de modèles pour la classification de tokens qui permettent de tokéniser les textes en propositions)

éd. Micha	éd. Sommer	<i>Lanzarote</i>	<i>Lancellotto</i>
et fet le signe de la <b>crois en mi son vis</b> puis embrace l escu	et met le signe de la <b>vraie crois en mi son front &amp; en mi son pis</b> puis embrace son escu	e fiço la señal dela <b>crúz sobre si</b> e abraço el escudo	e si fa el segno della <b>santa croce nel mi-luogo del suo viso</b> poscia imbraccia lo scudo
et <b>broche</b> le cheval des esperons	et <b>fiert</b> le cheual des esperons	e <b>firio</b> al cauallo de las espuelas	e <b>broca</b> il cavallo degli sproni

Extrait de la table d'alignement de certains des témoins

### Création d'une base lexicographique pour les langues anciennes et le français médiéval

The screenshot shows the entry for 'Cademoth' in the Biblissima+ interface. It includes the following sections:

- Declarations:** A table listing alternative lemmas (Lemme alternatif) for 'Cademoth', with a note 'Recensement des lemmes alternatifs' pointing to this section.
- Onomasticon Forcellini-Perin:** A table showing 'Cademoth' as an entry type with 0 references.
- Onomasticon De-Vit:** A table showing 'Cademoth' as an entry type with 0 references.
- Kirchenlateinisches Wörterbuch Sleumer:** A table showing 'Cademoth' as an alternative entry type with 0 references.
- Equivalence dans une autre langue:** A table showing the Greek equivalent 'Κεδομῶθ' (Keðomῶθ) with 0 references. A note 'Alignement avec des lexèmes équivalents dans d'autres langues' points to this section.
- Identifiants:** A table showing the lemma 'lilal.lemma' with the number 1829 and 0 references.

Alignement de dictionnaires

### Lemmatisation du français médiéval



À l'École des chartes avait été entraîné un **modèle d'étiquetage de l'ancien français**. Un des enjeux actuels est de convertir les référentiels utilisés par ce premier modèle afin de permettre l'interopérabilité des données en diachronie, ce qui nécessite notamment **l'alignement des lemmes du français médiéval**. Un autre enjeu est de poursuivre l'étiquetage de textes sur la période des XIV<sup>e</sup>-XV<sup>e</sup> siècles, afin de produire un modèle d'annotation permettant de couvrir l'ensemble des états de la langue médiévale.

### Deux corpus bibliques et patristiques



**BiBliIndex** rassemble les citations bibliques des textes chrétiens de l'Antiquité et du Moyen Âge, **Jerihna** les interprétations des noms hébreux issues de l'oeuvre de Jérôme de Stridon. Les alignements se font entre les référentiels de la Vulgate, de la Septante et de dictionnaires existants. Ils portent en premier lieu sur les noms propres des Bibles et leurs reprises dans les oeuvres de Philon d'Alexandrie, Origène, Eusèbe de Césarée et Jérôme.



### Un grand corpus de latin médiéval (700-1300)



La section de lexicographie de l'IRHT apporte un corpus lexical lemmatisé de 50 millions de mots de latin médiéval (800-1200), créé avec le soutien de l'ANR Velum. Biblissima+ permettra d'y ajouter 50 millions de mots supplémentaires (700 et 1300), pour l'étude de la diachronie à l'échelle Européenne et rendre compte du latin mérovingien et scolastique. Il sera interrogeable sur NoSketch-Engine et CQP-Web, consultable sous TEI-Publisher et téléchargeable au format XML.

### Deux lemmatiseurs (latin et grec)



**Collatinus** est un lemmatiseur et analyseur morphologique de textes latins (Philippe Verkerk).

**Eulexis** est un lemmatiseur de textes en grec ancien (Yves Ouvrard, Philippe Verkerk).

Tous deux font partie de la Boîte à outils Biblissima (Baobab).