

# Projet GRAPHEM

Grapheme-based Retrieval and Analysis for Paleographic Expertise on medieval Manuscripts (2008-2010)



Projet CNRS qui a pour ambition de contribuer à une paléographie objective assistée par analyse d'images

- Classification automatique (non supervisée) des écritures
- Moteur de recherche d'écritures similaires
- Moteur de recherche de mots

# Classification automatique

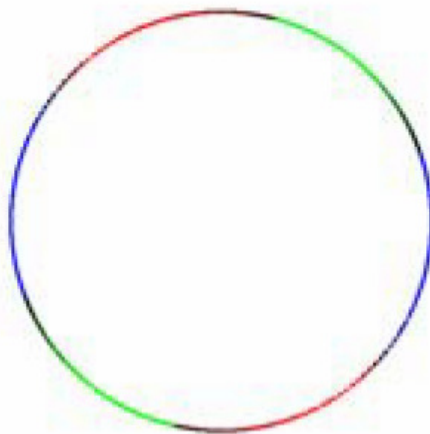
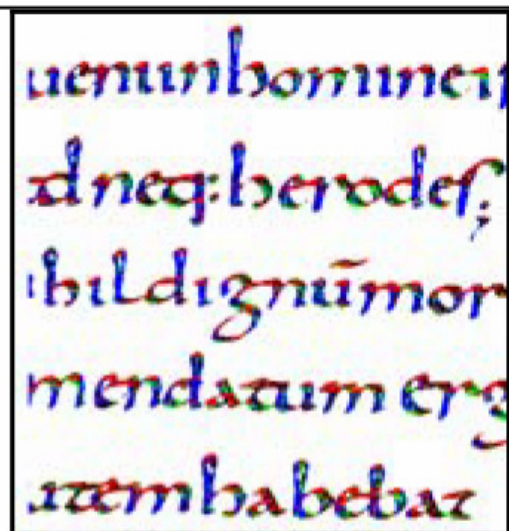
- Dépend des méthodes de classification
- Dépend des critères de la méthode
- Dépend des paramètres de la méthode
- Le nombre de classes est variable

Suivant la méthode et les paramètres nous obtenons entre 2 et 200 classes sur 4800 images

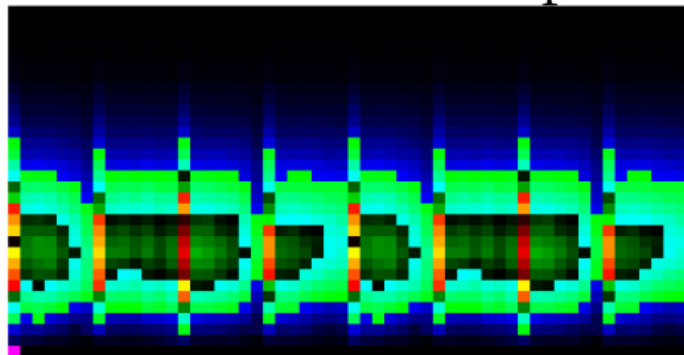
## Une classification

Trop fine → Moins d'erreurs mais trop de classes  
Trop grossière → Moins de classes mais des erreurs

# Classification automatique



Décomposition des traits selon les courbures orientées par Curvelets



Représentation compacte de sa signature

Des méthodes de classification globales, Indépendantes de la division en lettres.

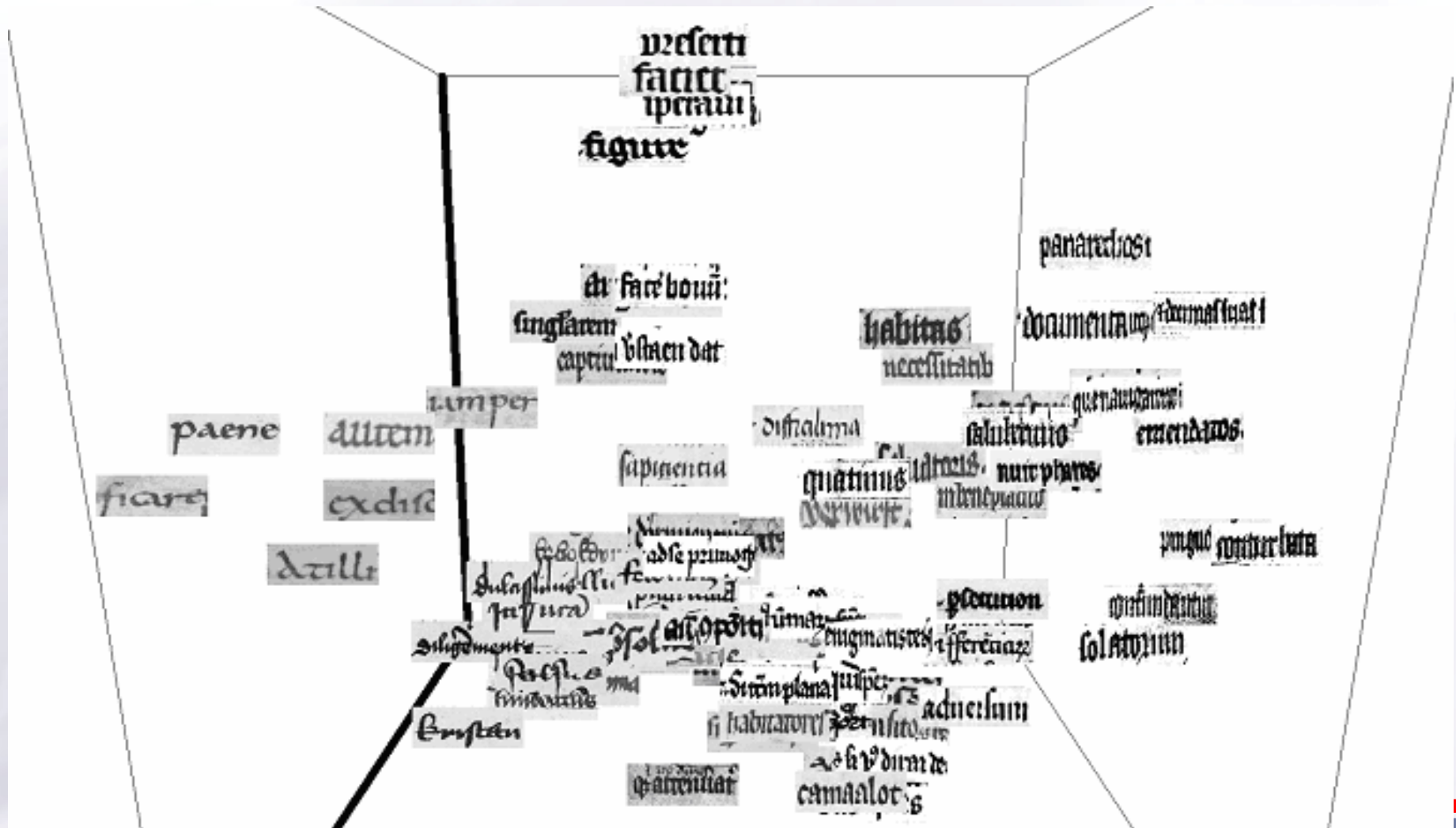
Exemple d'extraction de signature d'un échantillon.

# Classification automatique

30 images prises au hasard dans quelques classes

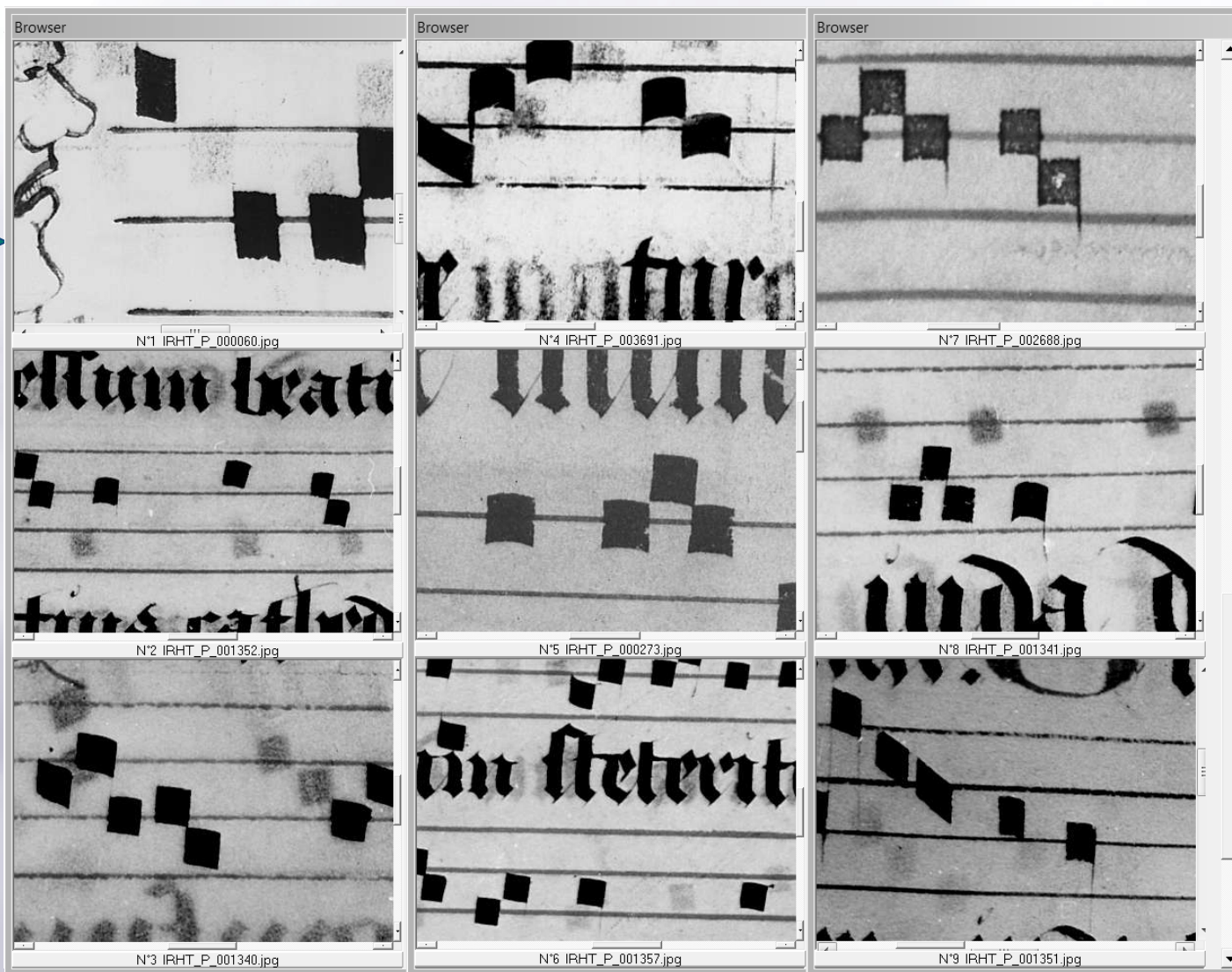


# Représentation multidimensionnelle



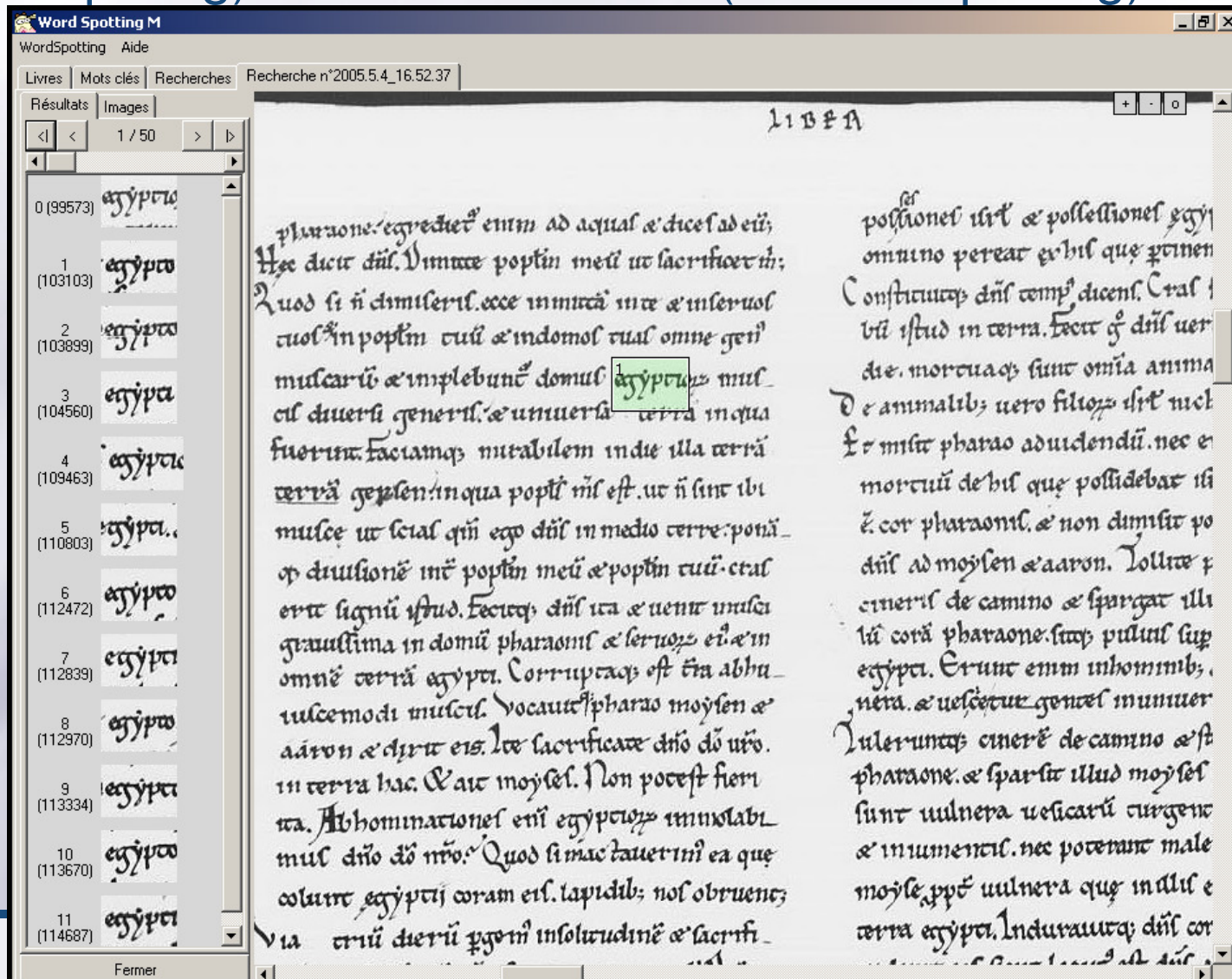
# Moteur de recherche d'écritures similaires

Requête

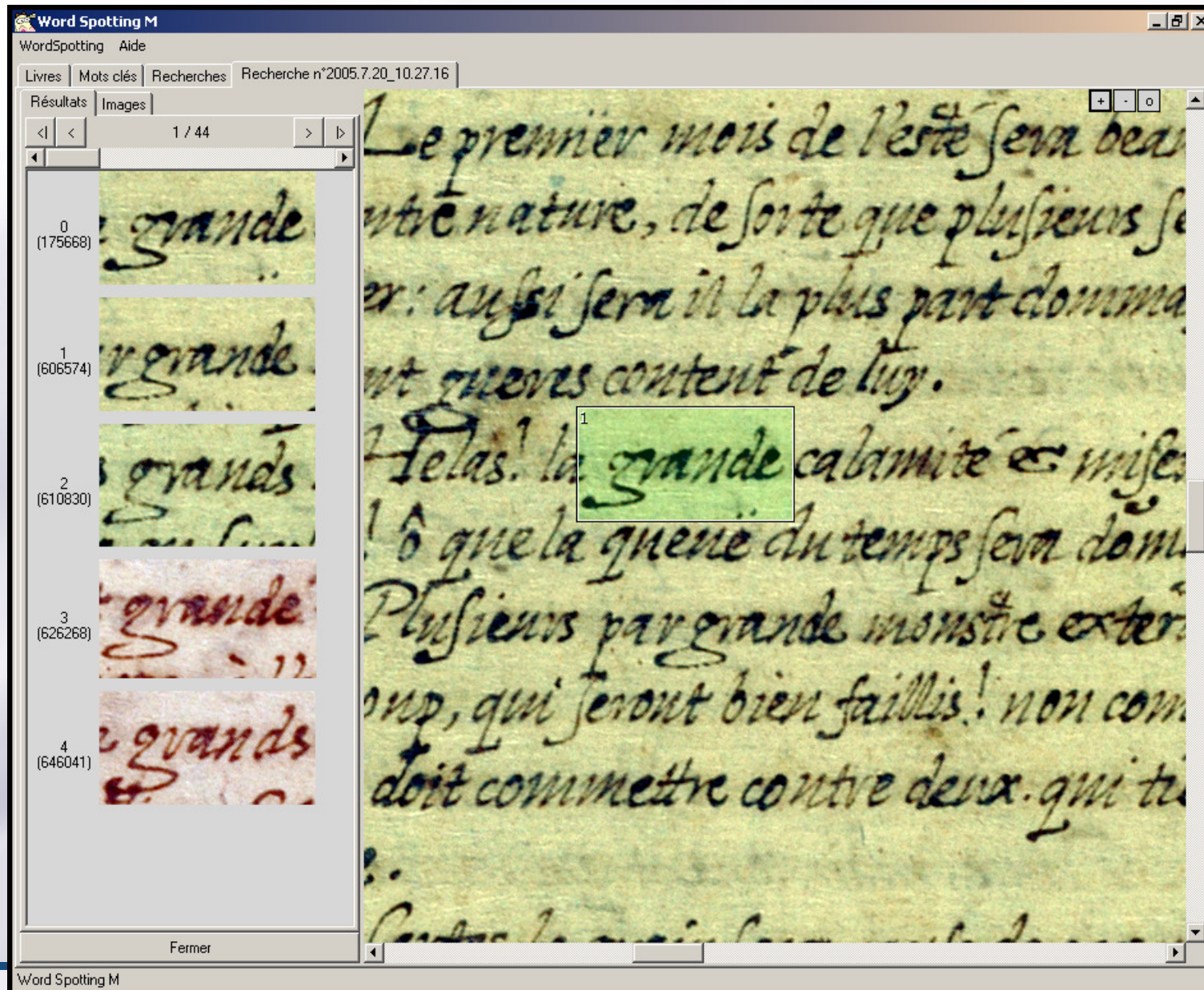


# Moteur de recherche d'images de mots

Trouve les images de mots similaires à une requête image (Word Spotting) ou dessin manuel (Sketch Spotting)



# Moteur de recherche de mots (wordspotting)



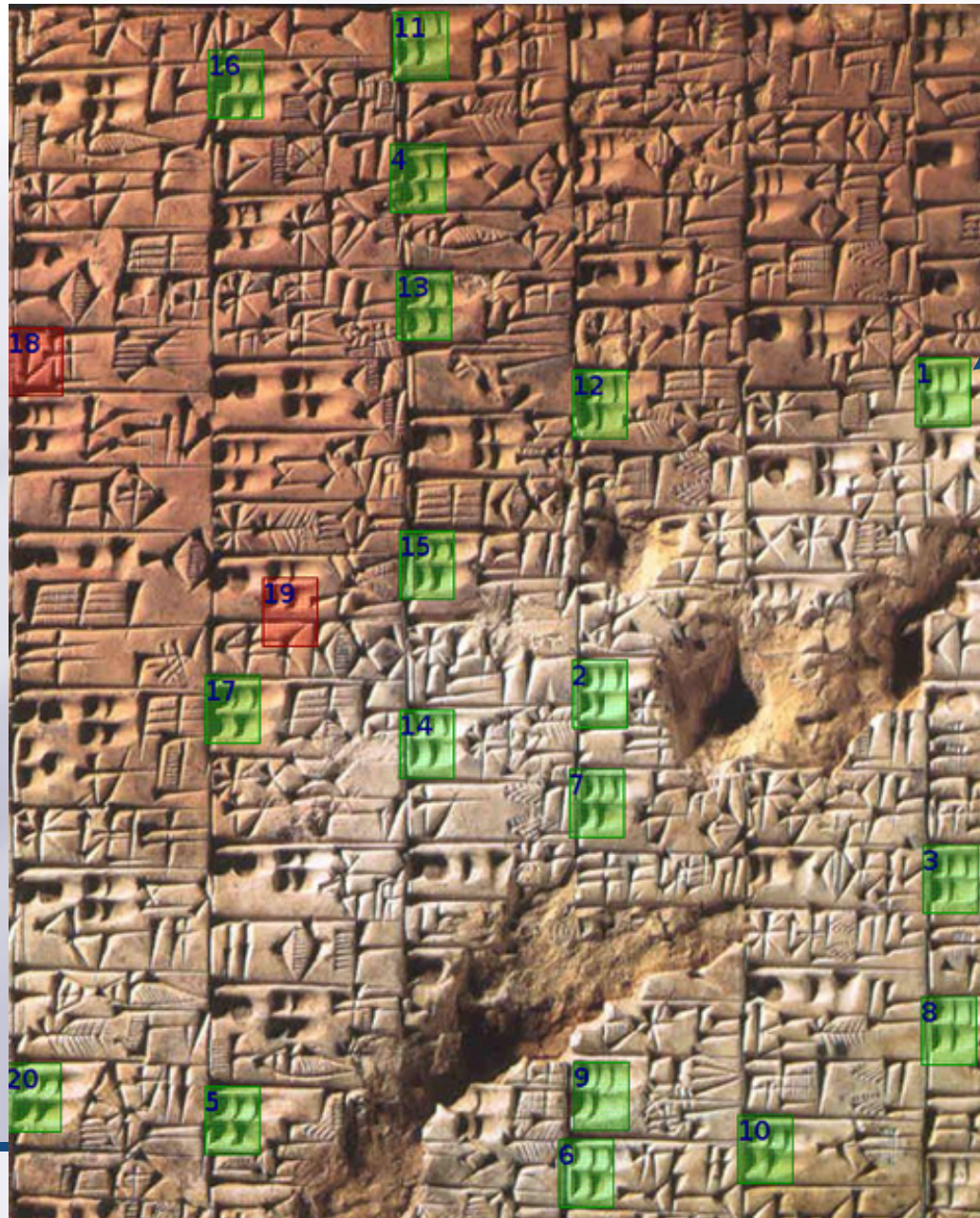


# Moteur de recherche de mots (wordspotting)

The screenshot displays the 'Word Spotting M' application window. The title bar reads 'Word Spotting M'. Below the title bar, there are tabs for 'Livres', 'Mots clés', and 'Recherches', with the current search identified as 'Recherche n°2005.9.23\_16.0.53'. The interface is divided into a left sidebar and a main viewing area. The sidebar, titled 'Résultats', shows a list of search results for the word 'محمد' (Muhammad), with each result accompanied by a small thumbnail image and a numerical ID (e.g., 0 (25292), 1 (25379), 2 (31033), 3 (31975), 4 (32199), 5 (32654), 6 (32885)). The main viewing area displays a large image of a manuscript page with Arabic calligraphy. A green rectangular box highlights the word 'محمد' in the second line of the text, with a small '1' above it. The text on the page includes: 'اللَّهُمَّ صَلِّ عَلَى سَيِّدِنَا مُحَمَّدٍ كَمَا صَلَّيْتَ عَلَى سَيِّدِنَا إِبْرَاهِيمَ ۖ وَبَارِكْ عَلَى سَيِّدِنَا مُحَمَّدٍ وَعَلَى آلِهِ سَيِّدِنَا مُحَمَّدٍ كَمَا بَارَكْتَ عَلَى آلِ إِبْرَاهِيمَ ۖ فِي الْعَالَمِينَ إِنَّكَ حَمِيدٌ مَجِيدٌ ۖ عَدَدَ خَلْقِكَ ۖ وَرِضَا نَفْسِكَ ۖ وَزِينَةَ عَرْشِكَ وَمِدَادَ كَلِمَاتِكَ ۖ وَعَدَدَ مَا ذَكَرَكَ بِهِ خَلْقِكَ ۖ فِيمَا مَضَى ۖ وَعَدَدَ مَا هُمْ ذَاكِرُونَكَ بِهِ ۖ فِيمَا بَقِيَ ۖ فِي كُلِّ سَنَةٍ وَشَهْرٍ وَجُمُعَةٍ وَيَوْمٍ وَلَيْلَةٍ وَسَاعَةٍ مِنَ السَّاعَاتِ'.

Traitement terminé.

# Moteur de recherche de motifs similaires



Requête

# Word retrieval

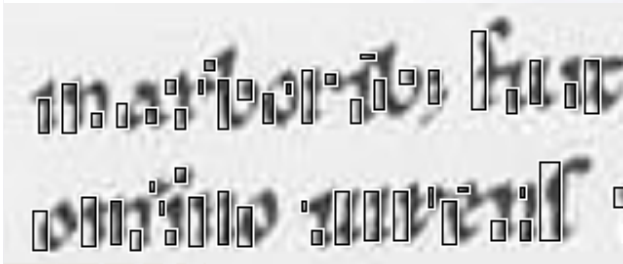
Moteur de recherche par une requête textuelle

« Google like query instead of query by image »

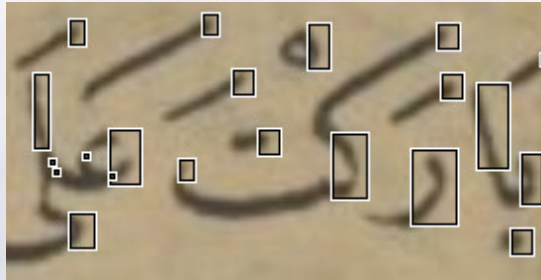
Saisissez le mot à rechercher.

a	b	c	d d	e e	f	g	h	i
k	l	m	n	o	p p	q	r	f f c s
t	u	x	y	z	et	?	pre	v v

# Word retrieval



Latin médiéval



Arabe



Chinois

## Repérage d'éléments guides

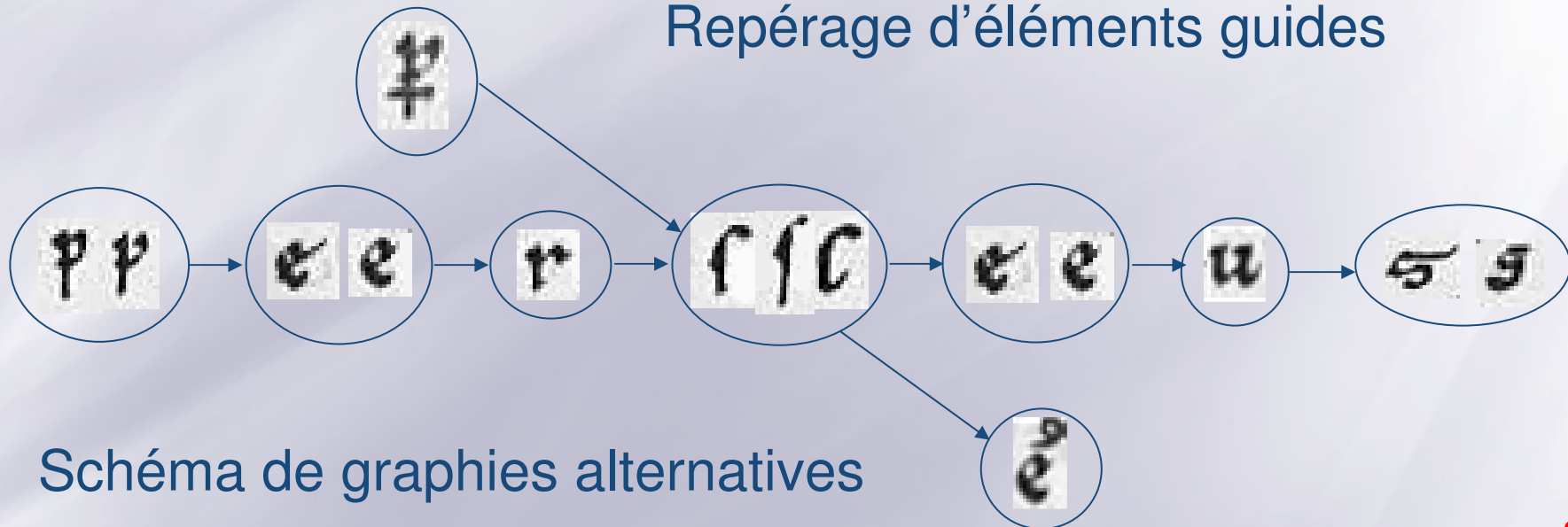


Schéma de graphies alternatives

# Word retrieval

Génération de graphies alternatives pour la recherche par similarité



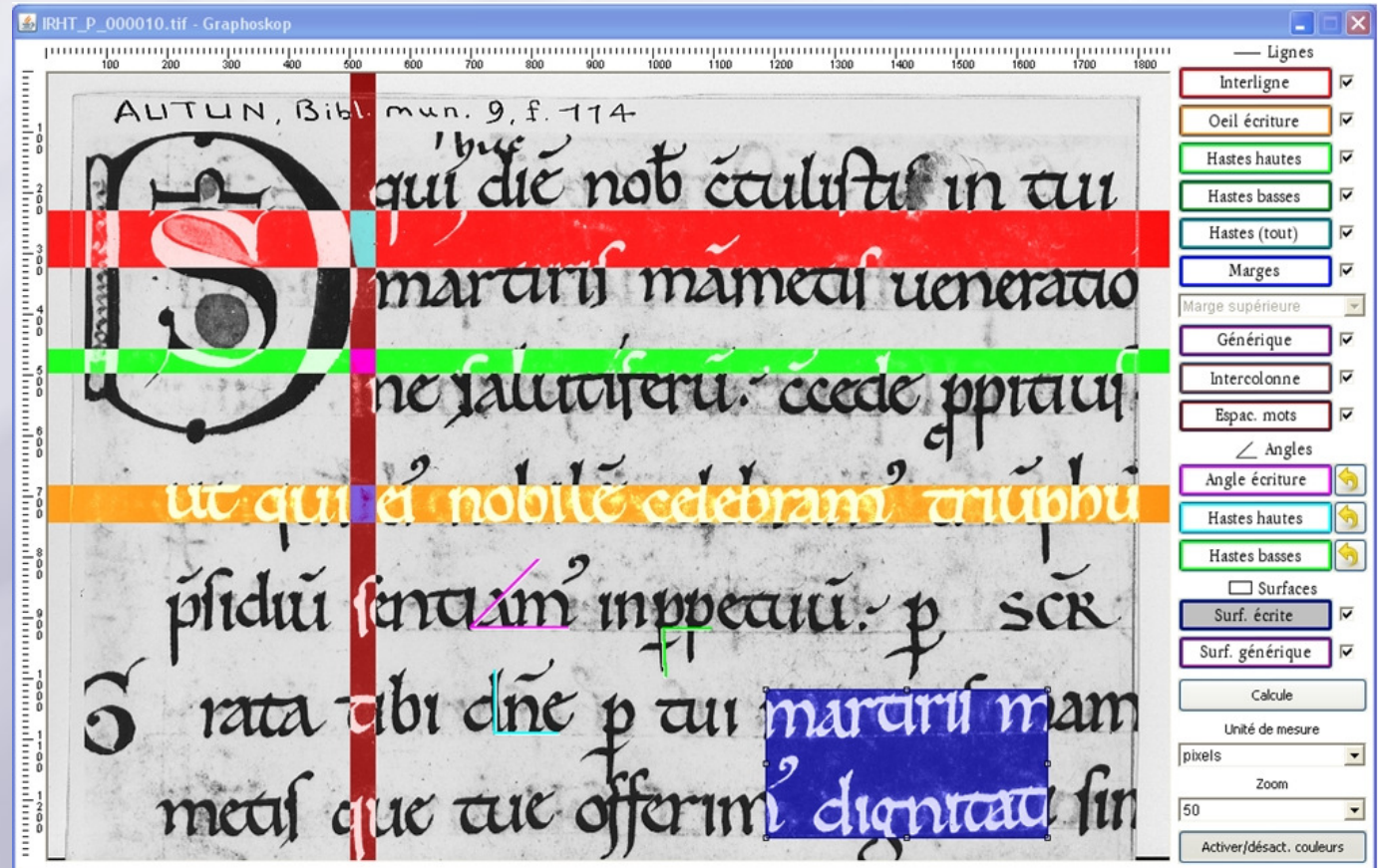
# Graphoskop

Plugin pour le logiciel d'imagerie biomédicale opensource ImageJ

Aide semi-automatisée à la mesure, après pose manuelle de repères :

- distances horiz. et vert.
- angles
- surfaces
- densité noir/blanc.

Définition de la nature des mesures par boutons (personnalisables dans la version 2)



# Graphoskop

Calculs statistiques  
en temps réel  
sur des mesures  
répétées

Mise à l'échelle

Exportation vers  
feuille de calcul

The screenshot displays the Graphoskop software interface. At the top left is the ImageJ window, and at the top right is the Results window showing a table of intervals in millimeters.

Intervalles (en mm)	
1	1.863
2	2.032
3	2.201
4	2.201

The main window shows a manuscript page with red horizontal lines indicating measurements. The right sidebar contains various analysis tools and settings:

- Lignes**
  - Interligne
  - Oeil écriture
  - Hastes hautes
  - Hastes basses
  - Hastes (tout)
  - Marges
  - Marge supérieure
  - Générique
  - Intercolonne
  - Espac. mots
- Angles**
  - Angle écriture
  - Hastes hautes
  - Hastes basses
- Surfaces**
  - Surf. écrite
  - Surf. générique
- Calculer
- Unité de mesure: mm
- Zoom: 50
- Activer/désact. couleurs

Web <http://liris.cnrs.fr/graphem/>



# Graphem

## Les Équipes

### Paléographie

ENC

IRHT

### Informatique

LIRIS

LIFO

CRIP 5

## Recherche

## GRAPHEM : Projet ANR

Le projet GRAPHEM est un projet pluridisciplinaire visant à l'analyse informatisée des écritures médiévales. Le programme a une durée de trois ans et est censé se conclure pour le 31 décembre 2010. Il a pour double ambition :

- de faire progresser la compréhension de l'évolution des formes de l'écriture,
- de créer des méthodes efficaces d'accès au contenu des manuscrits reposant sur la similarité de l'image des mots (Word-Spotting, Word-Retrieval).

La variété des écritures médiévales utilisant l'alphabet latin doit permettre d'élaborer et de tester des descripteurs de formes qui seront employés dans les deux cas. Une attention particulière sera portée à l'étude du graphème entendu comme l'élément minimal du tracé porteur d'une information pertinente.

Les laboratoires constituant le consortium sont le LIRIS, l'IRHT, le LIFO, le CRIP5 et l'Ecole nationale des chartes. Ils ont déjà travaillé ensemble, il y a quelques années, dans un projet d'exploration des manuscrits médiévaux, projet appartenant au programme "Société de l'information" du CNRS et intitulé "Formes et couleurs, outils de recherche". C'est dans ce contexte que la problématique de GRAPHEM a été conçue.

Les résultats escomptés sont de nature différente.

Pour la paléographie, il s'agit d'améliorer la typologie en mettant en jeu de nouveaux critères de discrimination des classes d'écriture.

Les méthodes d'accès au contenu textuel constituent une alternative aux méthodes de reconnaissance optique des caractères (O.C.R.), impuissantes sur les écritures anciennes. Elles ont vocation à être utilisées dans d'autres contextes que le manuscrit du Moyen Âge.

## Graphem

GRAPHEM : Projet ANR

Genèse

Objectifs

Liens

## Livrables

Rapports

Logiciels

## Editos

Annonce du projet dans la Gazette du livre médiéval

