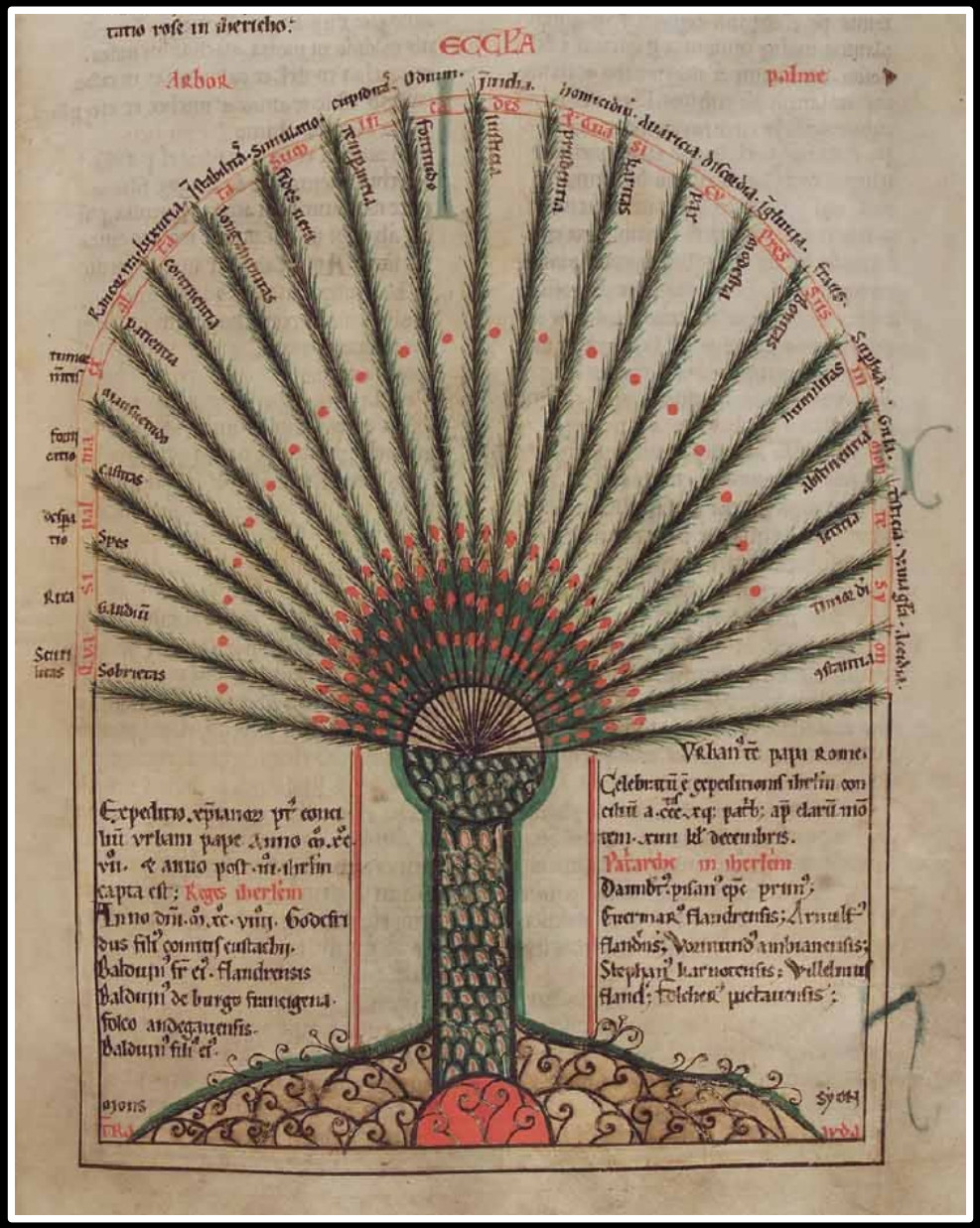
# FROM ACCUMULATION TO EXPLOITATION ?

## Experiments and proposals for indexing and for the use of diplomatics databases.




**Nicolas Perreaux** [UMR 5594 Artehis – Université de Bourgogne].

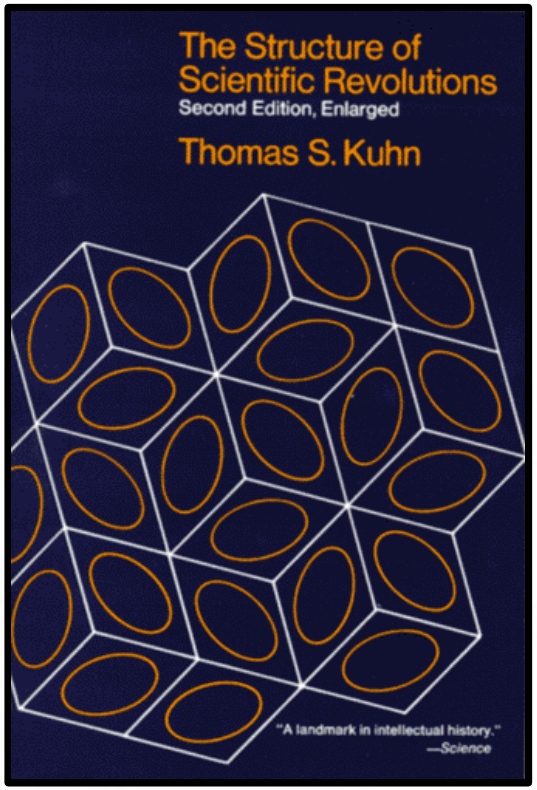
Cod. Guelf. 1 Gud. lat. (Lambert de Saint-Omer : *Liber floridus* – XII^e siècle), fol. 32r.

# Introduction


* **Discrepancy** between the **value** of charters databases, their **number** and their current **exploitation**.

* 1st **obstacle** : can traditional historical / diplomatics methods **manage so many documents** ?

➨ T.S. **Kuhn** : new tools = new paradigms ?




➨ Databases in medieval history =
a **double break**, **methological** but also **conceptual**.

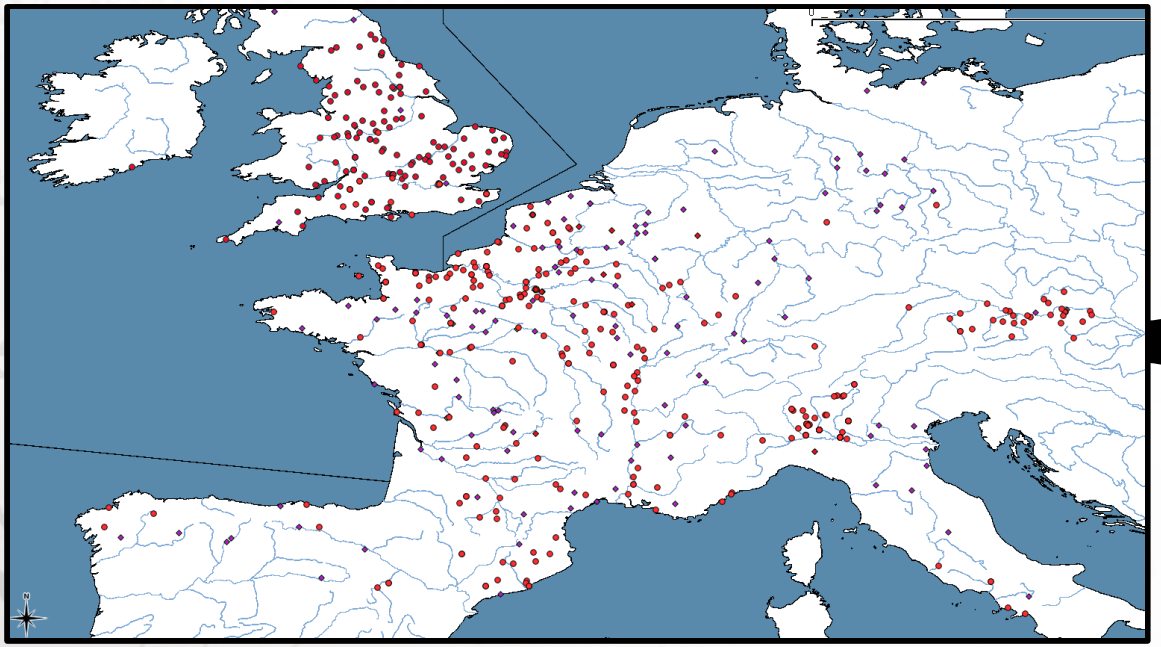
➨ **Data / Text-Mining** might be a way to get out this difficulty.

# I – *Corpora* or *corpus* ?

## 1. The creation of the database, the choice of a software

* Most of the **open** / available **charters** on the internet were collected.

+ **Help** of researchers + **Personal digitization** ≈> **150 000 charters** in total.



Charte Burgundie Medii Aevi

Telma. Chartes originales

DEEDS Project UNIVERSITY of TORONTO

FUNDACIÓ NOGUERA

inom Europe´s virtual documents archive

… + a lot more !

➨ It tooks 2 years to put everything in a single database (**XML**/TEI).

➨ **Philologic** : the only software that can handle +64k *corpora*.

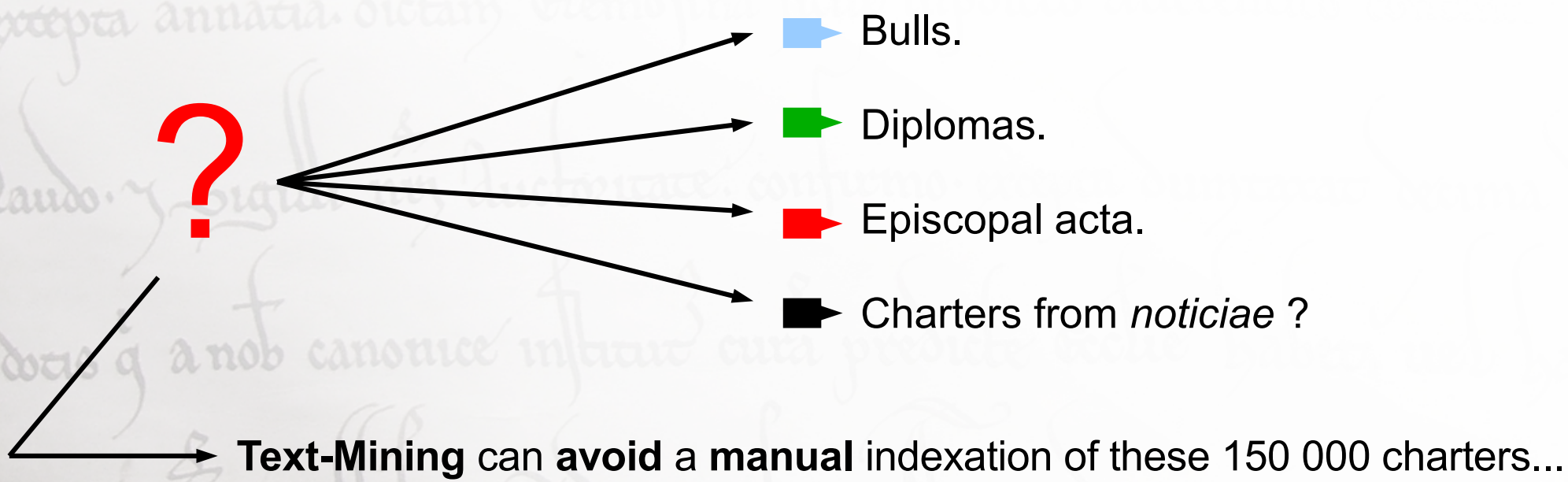
## I –
## 2. The need to automatically index the documents

* **Indexation** is a central criterion for a **proper exploration** of charters.

➨ Typological indexation **helps avoiding** a large number of « ***corpus* effects** ».

➨ Enables to **compare** the **vocabulary** of different types of charters, *etc.*

* Is it possible to distinguish automatically ?




**Text-Mining** can **avoid** a **manual** indexation of these 150 000 charters...

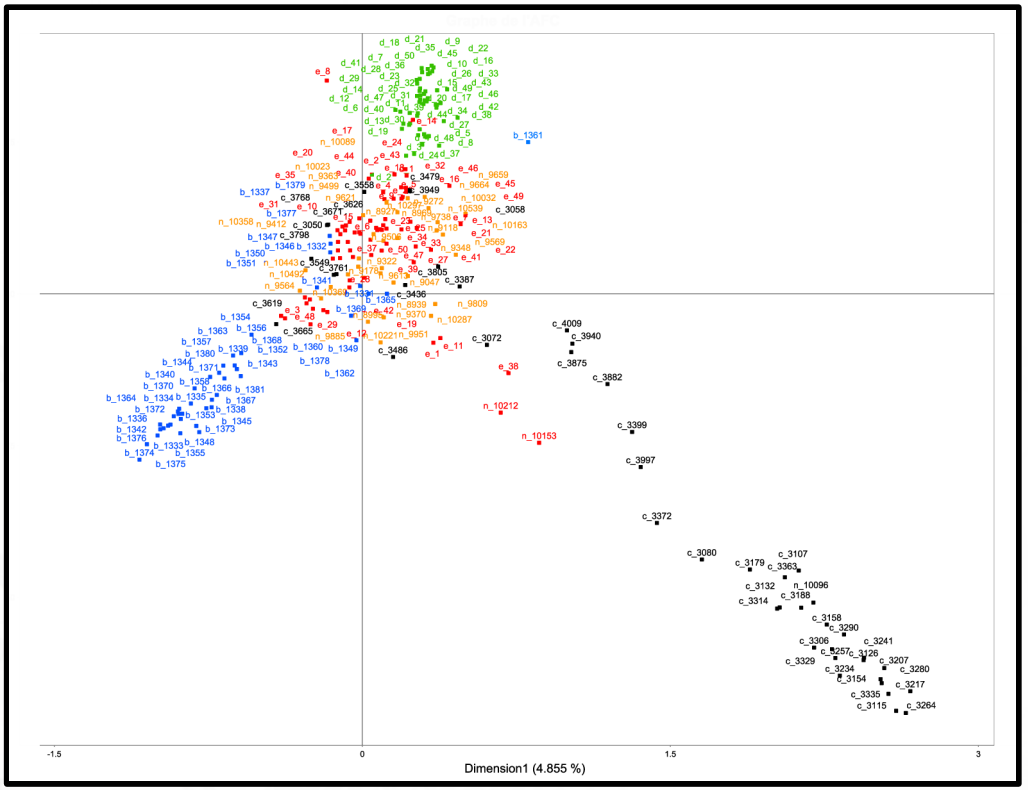
**I –**

## 2. *Measuring* the validity of the "traditional diplomatics categories" ?

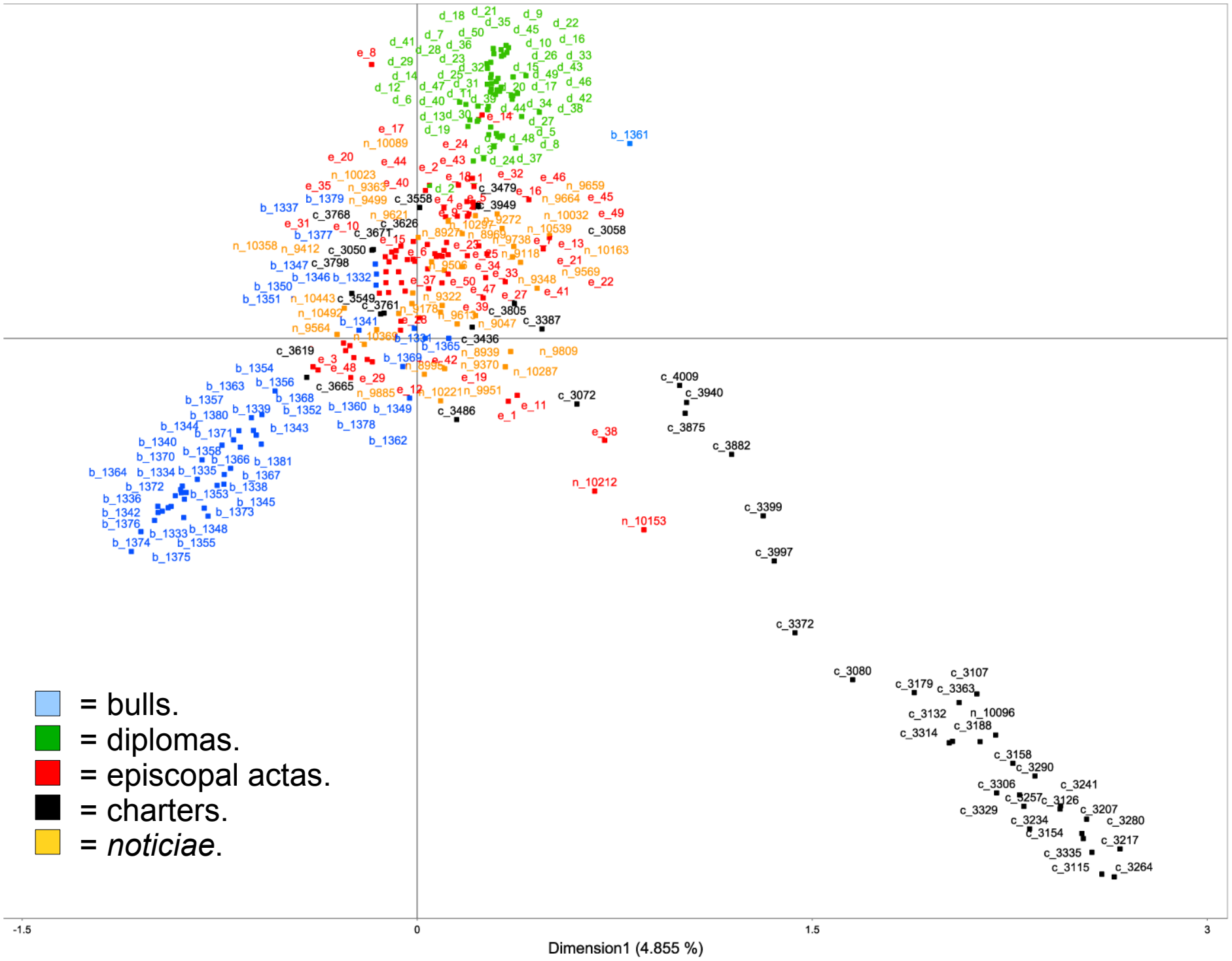
* Do **categories** in diplomatics cover a clearly **distinct vocabulary ?**

- Development of a **software** in order to measure the proximity of the vocabulary between charters (**Text-to-CSV**).
- Making of a **Factorial Analysis** on the output (*codage logique*)...

= bulls.
= diplomas.
= episcopal acta.
= charters.
= *noticiae*.



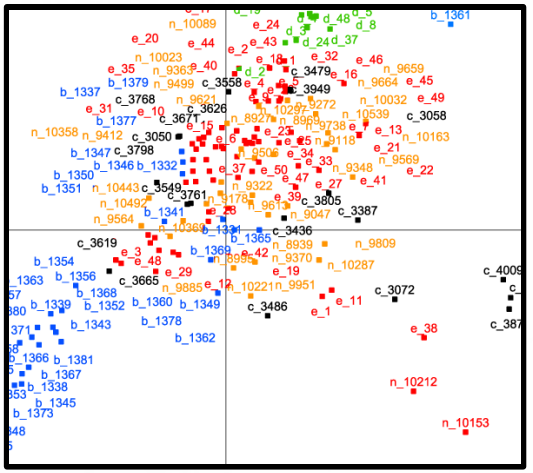

(factorial plan 1-2)


= bulls.
= diplomas.
= episcopal actas.
= charters.
= *noticiae*.
-1.5
0
1.5
3
Dimension1 (4.855 %)

**I –**

**2.**

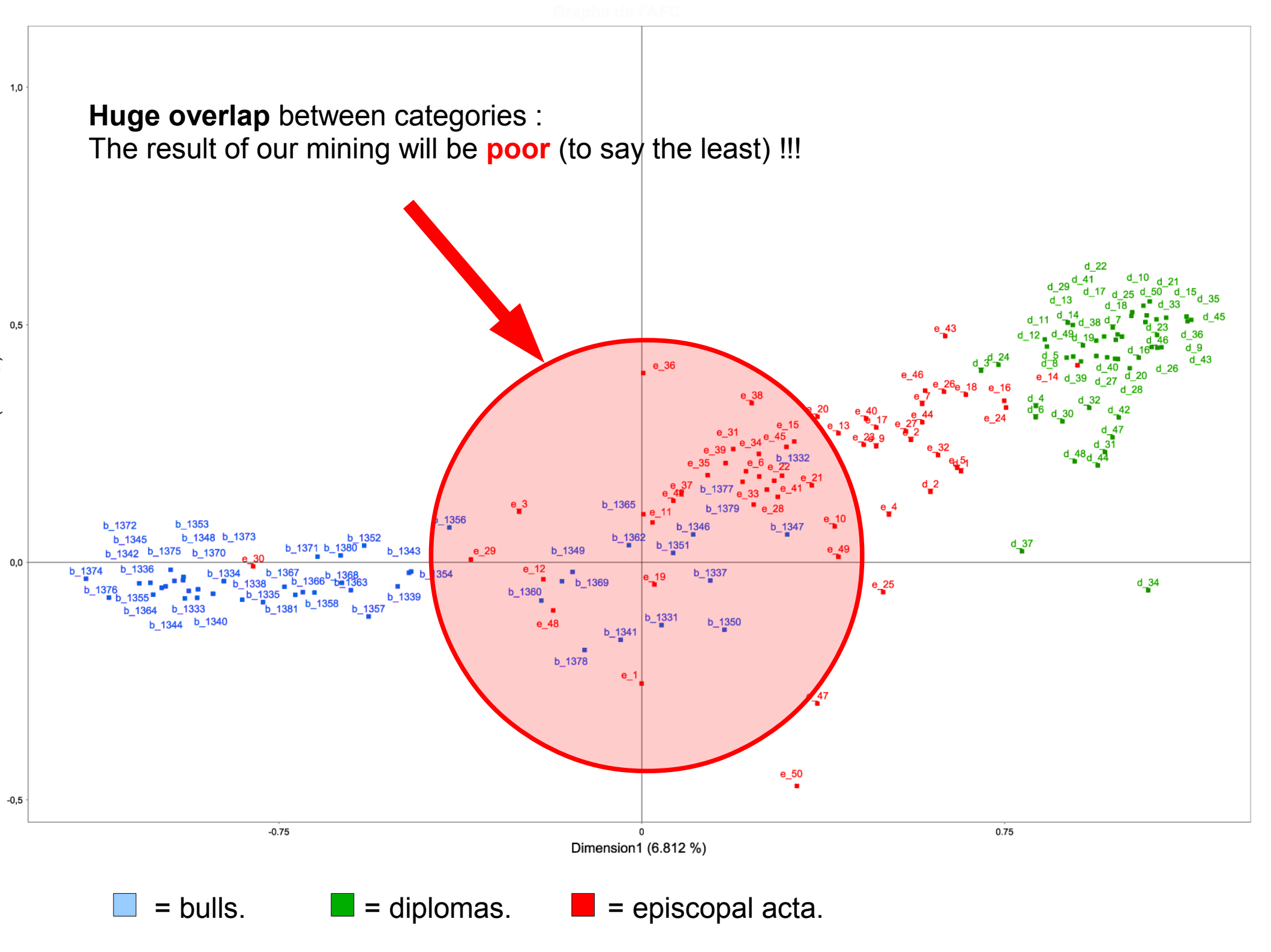
* Do **categories** in diplomatics cover a clearly **distinct vocabulary ?**

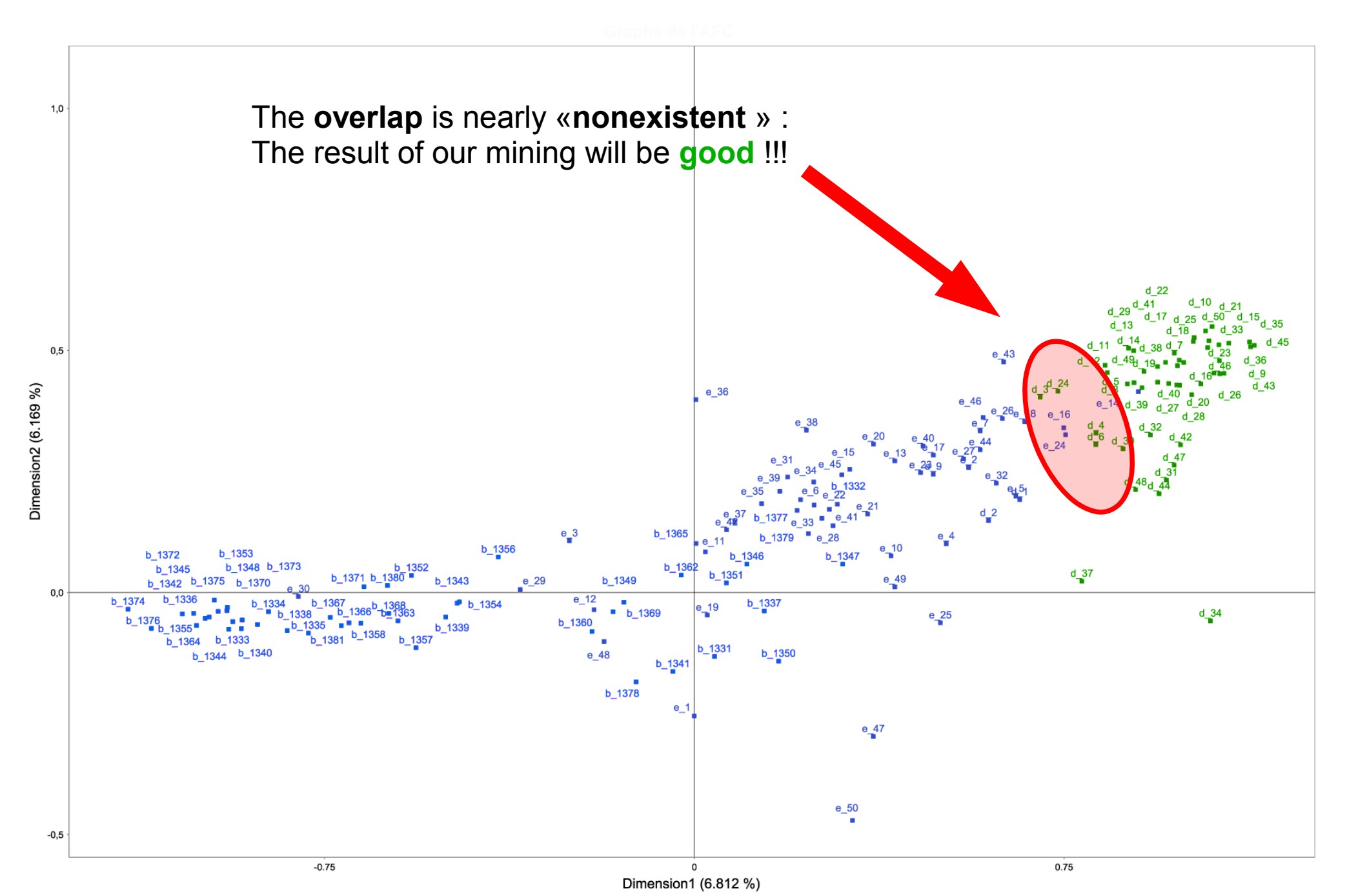


A test of **all categories at once** does not allow a good recognition (**overlap** between categories). TOO MUCH **NOISE** = **FAILURE** !

**Successive tests** on targeted categories = **SUCCES** !

**Example** : distinguishing **a.** Bulls. **b.** Diplomas. **c.** Episcopal acta ?


**Huge overlap** between categories :
The result of our mining will be **poor** (to say the least) !!!
1,0
0,5
0,0
-0,5
-0.75
0
0.75
Dimension1 (6.812 %)
= bulls.
= diplomas.
= episcopal acta.

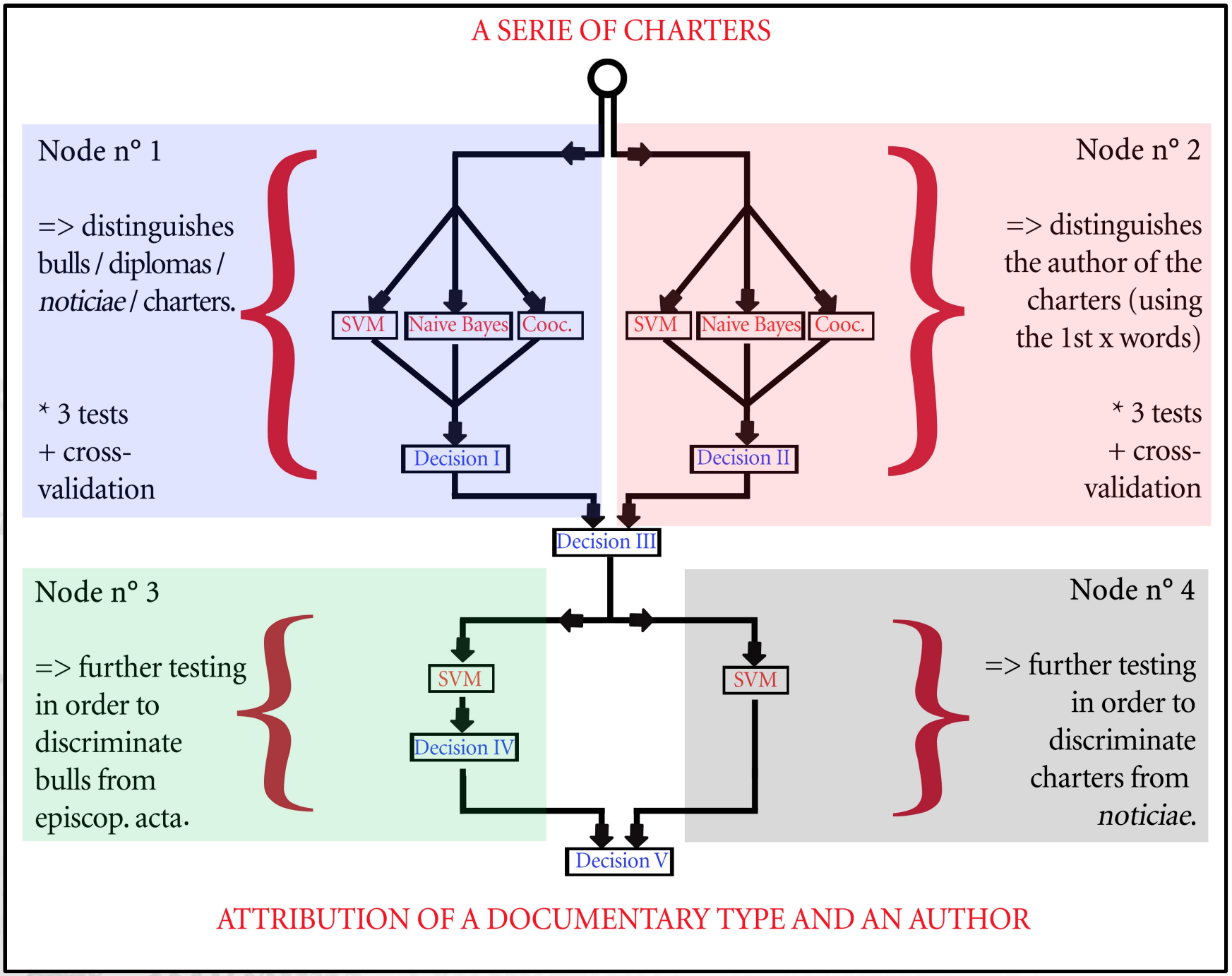



= bulls **+** episcopal acta.

= diplomas.

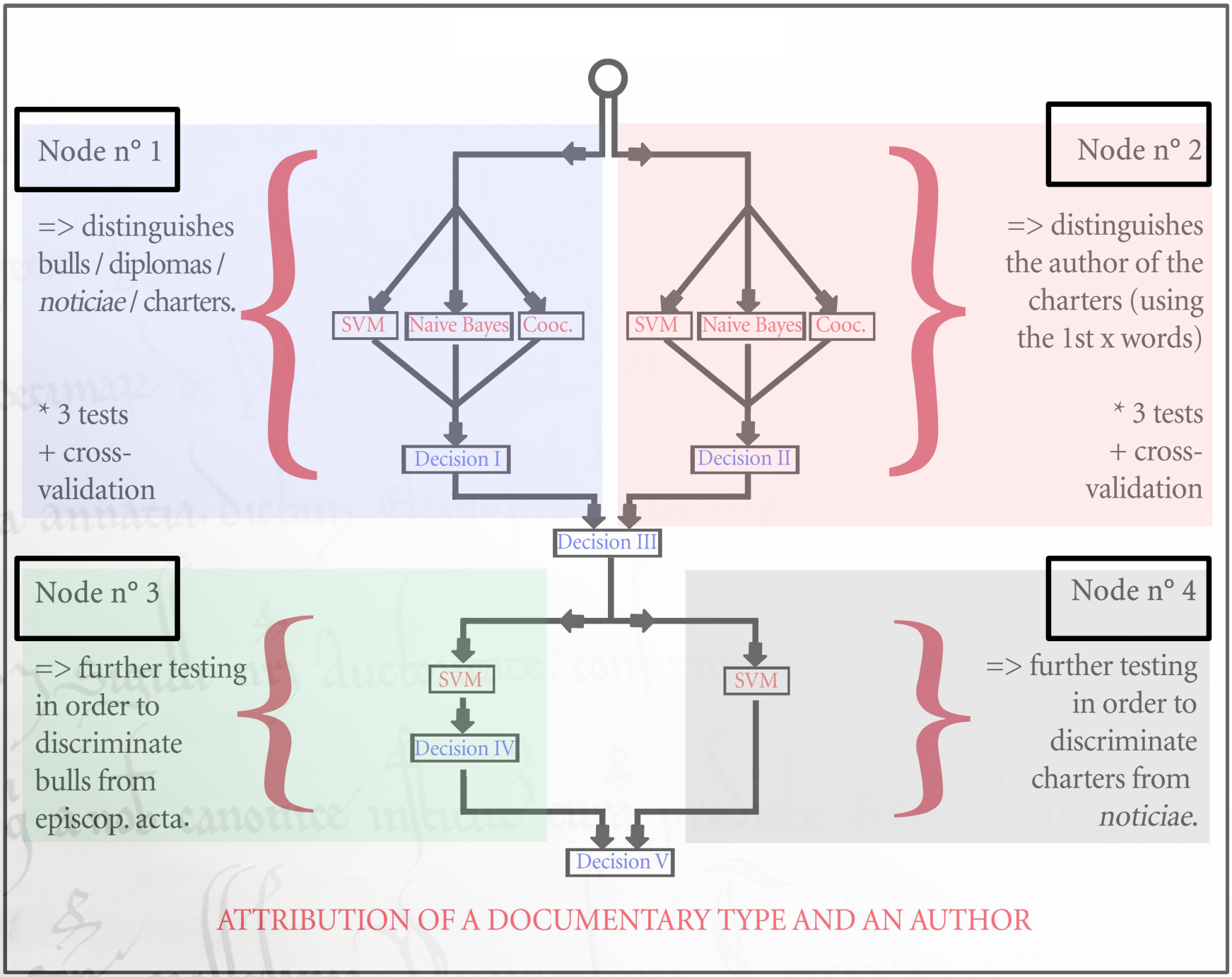
# II – The proposed algorithm for recognizing categories

## 1. Theoretical approach and model building



II –

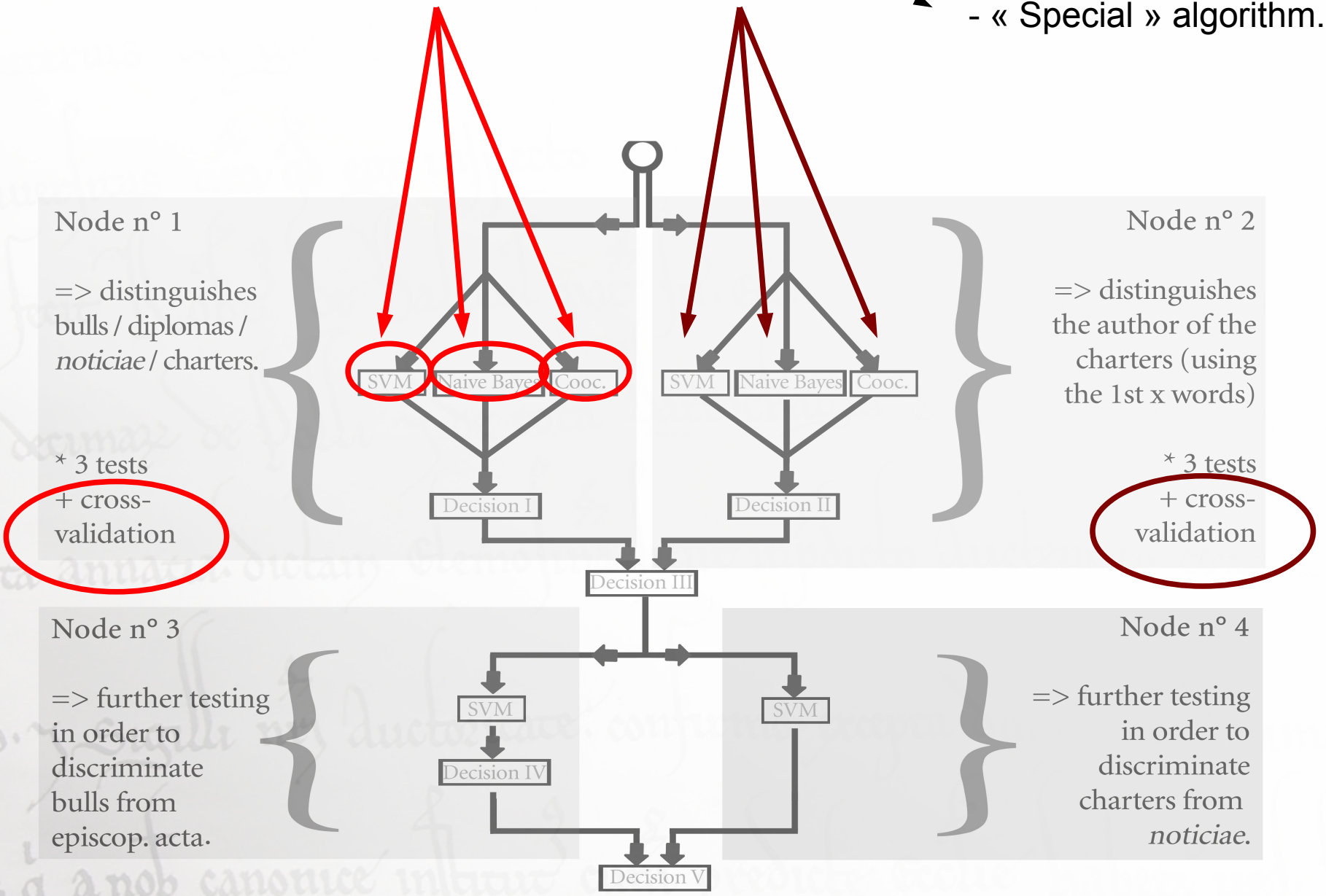
1.


Node n° 1
=> distinguishes bulls / diplomas / *noticiae* / charters.
* 3 tests + cross-validation
SVM
Naive Bayes
Cooc.
Decision I
SVM
Naive Bayes
Cooc.
Decision II
Node n° 2
=> distinguishes the author of the charters (using the 1st x words)
* 3 tests + cross-validation
Decision III
Node n° 3
=> further testing in order to discriminate bulls from episcop. acta.
SVM
Decision IV
SVM
Node n° 4
=> further testing in order to discriminate charters from *noticiae*.
Decision V
ATTRIBUTION OF A DOCUMENTARY TYPE AND AN AUTHOR

**II – 1.**

* **3 different algorithms** for the 1st two nodes

- **S**upport **V**ector **M**achine.
- **Naive Bayes**.
- « Special » algorithm.




ATTRIBUTION OF A DOCUMENTARY TYPE AND AN AUTHOR

* Results are directly integrated into Philologic.

=> **3 degrees** of reliability.

## II –
## 2. The validity of our method

* **Confusion matrix** = helps testing the results of our model.

➨ The **test** is, of course, made on **documents that are not present** in the "training database" (which now contains about 42,000 files).

* Improving the model = our main goal was to **reduce** the number of « **false positives** ».

* This method, **still in testing**, now automatically recognizes **for some regions** :

- ➨ 90% to 95% of the bulls.
- ➨ 90% to 95% of diplomas.
- ➨ 90% of episcopal acta.
- ➨ distinguishes 85% of noticia and 90% of the charters.

## II –

## 3. Complementary indexation : undated charters, chronological spans

* **Possible extension(s)** : Undated charters ? False documents ? etc.

➨ Seems to work quite well for the dating of undated documents (some **tests** have been done for the **cluniacs charters**... *work in progress*).

➨ The **problem** is then to **create a base of training files** for the institution / **region** from which the documents you want to date come from.

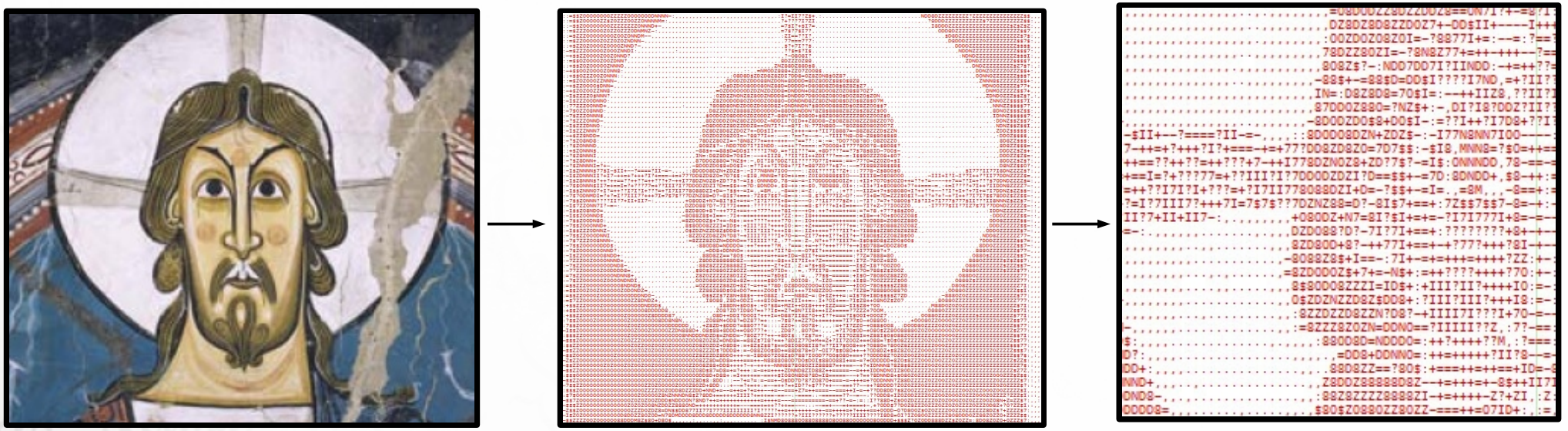
* **Last specificity** in our base : **Philologic** does not support time ranges (only one **single date *per* document**). Now :

➨ For each charter, **addition of two fields** : *terminus a quo*, *terminus ante quem* (we **changed** the **MySQL table loader**).

➨ New indexation that enable the **practical use of time spans**...

# III – Early experience(s) on our database

## 1. Presentation of Text-to-CSV



*Decomposing medieval documents ??? Text-to-CSV **do "the same thing"** to charters.*

- **Decomposing** cartularies / charters into **matrices**.
- Working on forms (**bag-of-words**) but also on larger parts of the *diplomatic discourse* : syntagms (**cooccurrences**).
- Manages several statistical **coefficients** (**TF-IDF**, etc.) and ***pruning***.
- **Clustering** is handled internally (algorithm by Mizuki Fujisawa).

→ The output files are directly usable under **R and Weka** !

## III –

## 2. Experience : writing charters, *formulae*, “zonation” [900-1050]

* Goal : **detect similarities** (and dissimilarities) between *corpora* **without making an *a priori* choice** on the vocabulary.

The adopted procedure (which was inspirated by) :

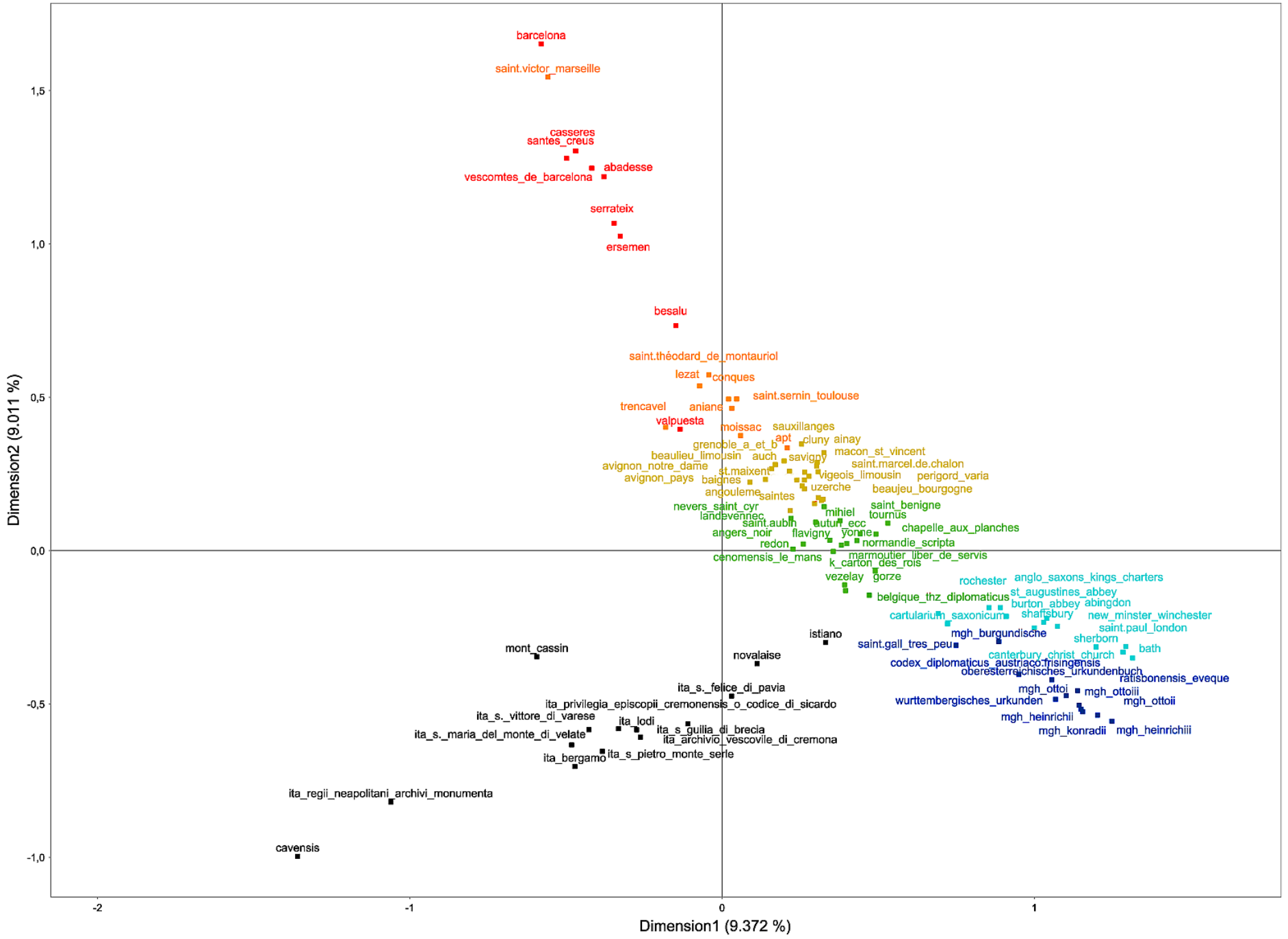
- The choice of a **time span** considered as more or less **homogeneous (900 to 1050)**.

- Test on **cooccurrences** : **3000** phrases, among the most frequent, were automatically retained.

- Creation of an **array** in “***codage logique***” (option included in Text-to-CSV).

- Use of **AFCs** (Factorial Analysis).
  (*This technique is now part of the Data-Mining “toolbox”*).

# III –
# 3. Result(s) and analysis



# Conclusion

1. **Vocabulary** of charters is **highly regionalized** in **large groups,** more or less **homogeneous**.

2. These **two experiments**, on indexing and regionalization must be seen as **a whole**.

3. A **better indexation** now goes through the identification of **areas** of the feudal system => key for **dating undated charters at large scale**, *etc.*

4. **Indexing**, **programming** are inseparable from the exploitation of the *copora*. This global process must be seen as **a whole**.

5. The perfect **software is a myth** : **medievalists themselves** should forge their own tools to get answer(s) to their specific questions.