Projet Tutoré

Charles Follet Roland Bary

12 mars 2015

Table des matières

I Numérisation

A l'aide d'un scanner standard, nous avons numérisé de la page 11 à 67 afin d'avoir le maximum de ressource pour notre travail suivant le 10 mars (date de remise de l'ouvrage).
Ensuite il a fallu convertir les fichiers .pdf scannés à l'aide d'un outil d'OCRisation en ligne de commande : tesseract. Nous avons obtenu des fichiers .txt exploitables pour créer des gazetiers.

II Annotations

Ébauches des règles d'annotation.

1 Spatiales

1.1 Villes, Départements, Pays

2 Temporelles

Les annotations temporelles concernent uniquement des périodes. Parmi ces périodes, on distingue celles d'émission des monnaies et celle de règne des souverains.

2.1 Périodes

■ **Vulgarisation :**

Date : trois et seulement trois chiffres. Peut être suivie d'un « / » et d'un chiffre traduisant l'incertitude sur la date.
Une période est un intervalle de deux dates. Dans notre travail, les dates sont constituées de trois et seulement trois chiffres. Chacune d'entre elle peut, en cas d'ambiguïté, être suivie d'un « / » et d'un chiffre traduisant l'indétermination de la période.
Exemple :
757/8

Période : intervalle de deux dates séparées par un tiret.
Exemple :
757/8-786

■

■ **Formalisation :**

Date

```
([0-9]{3}\/?[0-9]?)
```

Période

```
([0-9]{3}(\/[0-9])?)-([0-9]{3}(\/[0-9])?)
```

Exemple :
" *Type de 771-793/4 : Charlemagne (768-814),...*"
Group #1 : 771
Group #2 : 793/4

Group #1 : 768
Group #2 : 814

■

2.2 Périodes d'émission

■ **Vulgarisation :**

Périodes d'émission : « Type de {Période} : » (Période étant l'annotation définie précédemment). Il faut sécuriser la capture de la période pour ne pas récupérer toutes les périodes du document mais seulement celles correspondants à l'émission de monnaie en ajoutant la contrainte "précédée de Type de".

■

■ **Formalisation :**

Périodes d'émission :

```
Type de : ([0-9]{3}(\/[0-9])?)-([0-9]{3}(\/[0-9])?)
```

Exemple :

" *Type de 771-793/4 : Charlemagne (768-814)...*"
Group #1 : 771
Group #2 : 793/4

■

2.3 Périodes de règne

■ **Vulgarisation :**

Périodes de règne : « Nom_souverain {Période} : ». Il faut aussi sécuriser la capture grâce aux noms des souverains. Ceux-ci étant difficiles capter via une expression régulière, il faut créer un gazetier contenant tous les souverains. Ensuite, dès qu'une correspondance avec le gazetier sera établie on captera la période immédiatement après.

Exemple :

" *Type de 771-793/4 : Charlemagne (768-814)...*"

■

■ **Formalisation :**

Périodes de règne :

```
Nom_souverain ([0-9]{3}(\/[0-9])?)-([0-9]{3}(\/[0-9])?)
```

Exemple :

" *Type de 771-793/4 : Charlemagne (768-814)...*"
Correspondance : Charlemagne
Group #1 : 768

Group #2 : 814

3 Thématiques

3.1 Nature de la monnaie

■ **Vulgarisation :**

Nature de la monnaie : toujours un élément de l'ensemble Denier, Obole, Monnaie D'Or, Faux obole suivi ou non du nom d'un souverain. Il suffit donc d'utiliser une expression composé de tous les mots de l'ensemble. Nous vérifions qu'ils y a bien des espaces avant les termes recherchés afin d'augmenter la robustesse de notre recherche. Les noms des souverains seront trouvés à l'aide d'un gazetier.
Exemple :
"*Obole de Charles le Chauve*"

■

■ **Formalisation :**
Nature de la monnaie :

```
[\s]{2,}(Denier|Obole|Monnaie d'or|Faux Obole)(.*)? Nom_Souverain
```

Exemple :
"*Obole de Charles le Chauve*"
Correspondance : Charles le Chauve
Group #1 : 768
Group #2 : 814

■

3.2 Légende

■ **Vulgarisation :**

Légende : toujours à la ligne qui suit la nature de la pièce. Le revers droit est situé au début de cette ligne et commence par zéro ou un caractère +. Ensuite, vient une suite de 2 espaces ou plus. Pour finir, le revers droit vient se placer après zéro ou un caractère +.
Exemple :
"Denier de Charlemagne .....
+ CARLO 45ECROIX SIMPLE"

■

■ **Formalisation :**

Légende :

1. Se positionner à la ligne qui suit la nature de la pièce

   `(?:Denier|Obole|Monnaie d'or).*\n`

2. Capturer l'ensemble des caractères entre 0 ou 1 symbole + et 2 ou plus espaces. C'est la légende du droit.

   `\+?\s?(.*)[ ]{2,}`

3. Capture l'ensemble des caractères entre 0 ou 1 symbole + et la fin de ligne. C'est la légende du revers.

   `\+?\s?(.*)`

On obtient une expression comme suit :

```
(?:Denier|Obole|Monnaie d'or).*\n \s*\+?\s?(.*)[ ]{2,} \+?\s?(.*)
```

Exemple :
"Denier de Charlemagne .....
+ CARLO 45ECROIX SIMPLE"
Group #1 : CARLO 45E
Group #2 : CROIX SIMPLE

■

3.3 Types monétaire

■ **Vulgarisation :**

Le type monétaire est la simple concaténation de l'ensemble de mots "Type de" avec la période d'émission.
Exemple :
" Type de 771-793/4 : Charlemagne (768-814),..."

■

■ **Formalisation :**

```
(Type de [0-9]{3}\/?[0-9]?-[0-9]{3}\/?[0-9]?)
```

■

3.4 Ateliers

■ **Vulgarisation :**

Le catalogue est décomposé en ateliers, chaque début de "partie" ou "chapitre" est donc le nom de l'atelier. Ce nom correspond à un endroit géographique. Ce lieu peut être une ville, un lieu-dit, ... Il est difficile de trouver un pattern via les expressions régulières. Il faut constituer un gazetier.
■

3.5 Personnages

■ **Vulgarisation :**

Les personnages ont des formats aussi diverses que variés, il serait difficile d'utiliser une expression régulière. Il est plus judicieux d'utiliser un gazetier ici.
■

3.6 Collections, Trésors, Trouvailles

■ **Vulgarisation :**

Collections, Trésors, Trouvailles : sont chacun suivis de deux points. Ensuite vient le contenus concernant ces mots. Le contenu s'arrête lorsqu'on rencontre un point suivis d'un retour à la ligne ou bien d'un autre *mot* suivi de deux points.
■

■ **Formalisation :**

```
Mot_a_trouver:((?:.|\n)+?)(?:\.\s?\n|\w:)
```

Exemple :
"Collections : Berlin 1,77, 1,70, 1,59,
1,55 ; MEC 853 (1,78) ; Monnaie de Paris 105 (1,63) ; Prou 584 (1,58), 585 (1,69), 586 (1,79), 587
(1,72), 588 (1,77). Trésors :"
Group #1 : Berlin 1,77, 1,70, 1,59, 1,55 ; MEC 853 (1,78) ; Monnaie de Paris 105 (1,63) ; Prou 584 (1,58), 585 (1,69), 586 (1,79), 587 (1,72), 588 (1,77). Trésors :

■