Projet Tutoré

Charles Follet Roland Bary

11 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 gazetteers.

II Annotations

Ébauches des règles d'annotation.

1 Spatiales

1.1 Villes, Départements, Pays

Les villes ne sont pas très utiles à annoter, elles ont un sens thématique car elles correspondent à un atelier.

2 Temporelles

2.1 Périodes d'émission

■ **Cas général :**
Une période d'émission est un intervalle de deux dates. Dans notre travail, les dates sont constituée de trois et seulement trois chiffres. Chacune d'entre elles peut, en cas d'ambiguïté, être suivie d'un « / » et d'un chiffre traduisant l'indétermination de la période.
Définissons comment capter une date :

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

Signification : Trouver une suite de trois chiffres suivis de 0 ou 1 « / » suivis d'un chiffre. Placer le résultat dans un groupe.

Il suffit maintenant de combiner deux expressions régulières comme celle-ci avec un tiret et de concaténer l'expression obtenu à la chaine « Type de ».

```
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

■

2.2 Périodes de règne

■ **Cas général :**
Il est plus judicieux d'utiliser un gazetteer ici car les noms des personnages sont assez difficiles à capter. (L'OCRisation n'a pas été très réussie ... Solution ?)
■

2.3 ...

3 Thématiques

3.1 Nature de la monnaie

■ **Cas général :**

La nature de la monnaie est souvent un élément de l'ensemble Denier, Obole, Monnaie D'Or. 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.

```
[\s]{2,}(Denier|Obole|Monnaie d'or)
```

■

■ **Cas particulier :**

Il y a cependant des natures un peu "exotiques" comme faux obole.

■

3.2 Légendes

■ **Cas général :**

Construire un gazetteer concernant les légendes parait une tache périlleuse. De plus, étant donnée le grand nombre de légendes, la performance de notre application serai loin d'être optimale. Voila pourquoi nous proposons une expressions régulière, complexe certes, mais optimisée.

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 45E CROIX SIMPLE"
Group #1 : CARLO 45E
Group #2 : CROIX SIMPLE

■

3.3 Types monétaire

■ **Cas général :**
Le type monétaire correspond(?) à la période d'émission.

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

■

3.4 Ateliers

■ **Cas général :**
Le catalogue est décomposé en ateliers, chaque début de "partie" ou "chapitre" est donc le nom de l'atelier suivi. 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 gazetteer.
■
■ **Cas particulier :**

■

3.5 Personnages

■ **Cas général :**
Les personnages ont des formats aussi diverses que variés, il serait difficile d'utiliser une expression régulière. Il est plus judicieux d'utiliser un gazetteer ici. (L'OCRisation n'a pas été très réussie ... Solution ?)
■
■ **Cas particulier :**

■

3.6 Collections, Trésors, Trouvailles

■ **Cas général :**
Les *mots* 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.

```
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 :

■

■ **Cas particulier :**

■