Projet Tutoré

M1 Technologies de l'Internet

Conception et développement d'une application d'annotation thématique dans l'environnement Gate

---

Auteurs :
Roland Bary
Charles Follet

Tuteurs :
Marie-Noëlle Bessagnet
Annig Lacayrelle
Albert Royer
Christian Sallaberry

Remerciements

Nous tenons à remercier nos tuteurs pour leur pédagogie et leur encadrement. Monsieur Royer pour sa précision et sa connaissance pointue du domaine. Madame Lacayrelle pour son soutien et sa clarté. Monsieur Sallaberry pour nous avoir remis sur de bonnes pistes quand nous nous égarions. Et enfin, madame Bessagnet pour avoir assuré la coordination et le suivi de ce projet.

Table des matières

Chapitre 1

Introduction

La production de documents se fait de plus en plus rapidement et facilement grâce la multiplication des moyens numériques mis à notre disposition. Aujourd'hui, chacun peut écrire ce qu'il veut, quand il le veut et le publier quasi instantanément.
Cette liberté est une avancée certaine dans le domaine de la communication. Mais, alors qu'il est possible de chercher des mots via notre éditeur de texte préféré comment faire des recherches un peu plus avancées dans les divers documents ?
Plus précisément, posons la problématique suivante :

"Comment effectuer des recherches avancées sur un corpus de document non structuré ?"

C'est dans le cadre de notre projet tutoré que nous allons, sans avoir la prétention de le résoudre, travailler sur ce problème. Le problème est, dans notre cas, limité à un thème bien précis : les monnaies carolingiennes.
Notre démarche s'appuiera des connaissances actuelles que nous présenterons en première partie. Ensuite, viendra notre principe de résolution du problème.

Chapitre 2

État de l'art

Dans cette partie sera décrit les quelques connaissances et outils actuels dans le domaine considéré : l'annotation sémantique.

I Connaissances

II Outils

Chapitre 3

Résolution

Introduction

Dans cette partie sera décrit notre démarche pour résoudre la problématique. Voici une version résumée de notre raisonnement pour venir à bout de ce projet. Elle donne une idée générale de notre démarche. Nous avons travaillé par raffinement successif.

1. Amélioration de notre connaissance sur le sujet.
2. Description en langage naturel des règles d'extraction d'information.
3. Formalisation et amélioration des règles précédemment établies.
4. Traduction des règles formelles en langage JAPE.
5. Construction de gazetiers pour des règles non définissables via JAPE.

Nous allons d'abord présenter lister l'ensemble des règles à annoter avec la solution choisie (JAPE ou gazetier). Ensuite nous décrirons précisément chaque règle définie avec JAPE et chaque règle définie avec un gazetier.

I Amélioration de nos connaissances

// Avec le livre

II Informations à annoter

Nom de la règle	Solution choisie
Spatiale	
Villes, Départements, Pays	?
Temporelle	
Période d'émission	JAPE
Période de règne	JAPE
Thématique	
Natures de la monnaie	JAPE
Légendes	JAPE
Types monétaire	JAPE
Collections, trésors, trouvailles	JAPE
Ateliers	Gazetier
Personnages	Gazetier

III Annotation avec JAPE

Périodes

■ *Vulgarisation :*

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

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

■
■ *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

```
    // Regle JAPE
  Macro: TROIS_NOMBRES
({Token.kind==number,Token.length == 3})

Macro: UN_NOMBRE
({Token.kind==number,Token.length == 1})

Macro:SLASH
```

```
({Token.string=="/"})

Macro:DATE_PRECISE
(TROIS_NOMBRES)

Macro:DATE_IMPRECISE
(TROIS_NOMBRES SLASH UN_NOMBRE)

Macro:DATE
(DATE_PRECISE | DATE_IMPRECISE)

Rule: PeriodeRule
(
   (DATE):d1({Token.string =="-"})(DATE):d2
    ):Periode -->
:Periode{/*Code java pour extraire les extremites de l'intervalle*/}
```

■

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

```
   // Regle JAPE
   Macro: CHAINE_DEBUT
(
    ({Token.string =="Type"})({SpaceToken})
    ({Token.string =="de"})({SpaceToken})
)

Rule: PeriodeEmissionRule
(
    CHAINE_DEBUT ({Periode}):p
):PeriodeEmission
-->
:PeriodeEmission.PeriodeEmission = { Kind = "PeriodeEmission" ,D1 =
    :p.Periode.D1, D2 = :p.Periode.D2}
```

■

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

■

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

■

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

■

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]?)
```

■

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 :

■

IV Annotation avec Gazetiers

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.
■

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.
■

Chapitre 4

Conclusion