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

Première partie

Introduction

Avec l'évolution de manière significative des volumes d'informations sur internet, on peut observer une évolution du web vers une approche dans laquelle chaque donnée acquiert un sens afin de rendre possible une interprétation du contenu des pages web par des machines. Cette extension constitue le web sémantique.L'une des principales motivations du web sémantique est la recherche d'information sémantique.
C'est donc dans ce cadre que nous sommes intervenus pour répondre à l'appel d'offre de nos encadrants. L'objectif est l'annotation sémantique d'un document texte spécifique, qui constitue effectivement la première étape dans un processus d'indexation et de recherche d'information sémantique.

Au regard de ce qui a été exprimé en amont, se pose les problématiques suivantes :

- Existe-t-il des outils qui se prêtent aisément à l'annotation sémantique ?
- Quelle approche de conception peut nous permettre de réaliser cette étape d'annotation sur un document texte non-structuré ?

La résolution de ces différentes problématiques, nous à donc amené à organiser ce document comme suit : /*Il faut caller notre plan ici */ :) Une première partie dans laquelle nous présenterons le cahier des charges. Ensuite une seconde partie décrira quelques connaissances existantes sur le sujet avec les technologies utilisées au sein du projet.

Deuxième partie

Cahier des charges

I Contexte

A partir des travaux de Georges DEPEYROT sur les monnaies carolingiennes, nous avons travaillé pour une équipe parisienne de numismates sur l'annotation du Numéraire Carolingien [1].
Sur celui-ci, l'équipe a besoin d'effectuer des recherches :

Temporelles : Quelles étaient les pièces en circulation de l'an 859 à l'an 865 ?

Spatiales : Dans quels ateliers, les pièces de type Obole de Charlemagne ont été produites ?

Thématiques : Combien d'exemplaires de la monnaie d'or de Charles le Chauve ont été étudiés ?

Répondre à cette demande implique de définir puis d'explorer les dimensions temporelles, spatiales et thématiques de l'ouvrage.
Pour cela, il est nécessaire de connaître le domaine et l'ouvrage afin de savoir quelle information correspond à quelle dimension.
Une fois cet apprentissage fait, nous pouvons construire des règles dans une chaine de traitement permettant d'annoter chaque information en fonction de sa dimension.

Les monnaies carolingiennes sont le domaine central pour la réalisation du projet. Les ressources nécessaires à l'annotation (ici sous forme de gazetiers) ont été construites à partir des données de l'ouvrage.

Forte de son expérience dans le domaine, la maîtrise d'ouvrage nous a demandé d'utiliser la boîte à outils logicielle GATE qui sera utile pour le traitement du langage naturel.

En résumé, les caractéristiques du projet sont :
- l'apprentissage et la compréhension du domaine considéré,
- l'étude des principes d'annotation de documents,
- le développement d'une chaîne d'annotation dans GATE,
- la mise en place d'une visualisation des résultats.

1. `http://www.cgb.fr/le-numeraire-carolingien-moneta-77-3e-edition-depeyrot-georges,Ln71,a.html`

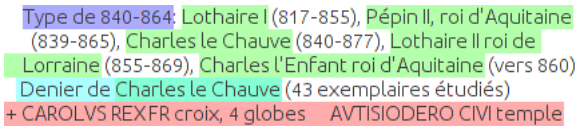
II Description

La chaine de traitement prend un document textuel en entrée, produit un document XML en sortie et le met en forme pour une meilleure lisibilité.
Exemple :

Illustrons par un scénario les objectifs de la chaine de traitement. Elle prend par exemple en entrée le texte suivant

Type de 840-864: Lothaire I (817-855), Pépin II, roi d'Aquitaine (839-865), Charles le Chauve (840-877), Lothaire II roi de Lorraine (855-869), Charles l'Enfant roi d'Aquitaine (vers 860) Denier de Charles le Chauve (43 exemplaires étudiés)
+ CAROLVS REX FR croix, 4 globes AVTISIODERO CIVI temple

et l'annote



Chaque information pertinente est annotée. En bleu la période d'émission de la monnaie (Temporel), en vert les souverains qui l'ont faite produire (Thématique), en cyan la nature de la monnaie (Thématique) et en rouge la légende (Thématique).
Afin de développer cette chaîne,nous avons du planifier notre travail et nos réunions avec la maîtrise d'ouvrage. Cette planification sera présentée dans la partie suivante.

III Diagramme de Gantt prévisionnel

Troisième partie

Cadre d'analyse

Quatrième partie

Développement

Introduction

VI Prise en main de l'environnement

VII Définition des dimensions d'annotation et leur contenu

VIII Première recherche d'entités nommées avec les gazetiers

Ateliers

■ *Vulgarisation :*

L'ouvrage est décomposé en ateliers, chaque début de "partie" ou "chapitre" est donc le nom de l'atelier. Ce nom correspond à un endroit géographique de France ou de ses pays limitrophes. Ce lieu peut être une ville, un lieu-dit, ... Il peut porter un nom qui n'existe plus aujourd'hui. Pour cela, il est difficile de trouver un pattern via les expressions régulières. Nous avons construit ce gazetier grâce partir du début de l'ouvrage qui recense tous les ateliers.
■

■ *Formalisation :*

```
Aix-la-Chapelle (Allemagne)
Agen (Lot-et-Garonne)
Aix-la-Chapelle
Alsheim (Allemagne)
Altenheim (Bas-Rhin)
Amiens (Somme)
```

■

Souverains

■ *Vulgarisation :*

Les souverains ont des formats aussi diverses que variés.Ils comportent des majuscules, des chiffres romains... Il serait difficile d'utiliser une expression régulière pour espérer annoter cette information. Il est plus judicieux d'utiliser un gazetier. Il sera construit à partir du début du numéraire.
■

■ *Formalisation :*

```
Pépin le Bref:valeur=Pépin le Bref:periode=752-768
Adalbert Lothaire:valeur=Adalbert Lothaire:periode=954-986
Amoul roi de Germanie:valeur=Amoul roi de Germanie:periode=887-899
Bérenger I:valeur=Bérenger I:periode=888-924
Bérenger II:valeur=Bérenger II:periode=950-961
```

■

IX Deuxième recherche d'entités nommées avec les règles JAPE

Périodes

■ *Vulgarisation :*

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

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

■

■ *Règle JAPE :*

```
     // 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 :*

Une période d'émission à la forme suivante : « Type de {Période} : » (Période étant l'annotation définie précédemment).
Il faut sécuriser la capture de la période d'émission 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 :*

`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

■

■ *Règle JAPE :*

```
   // 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 :*

Une période de règne à la forme suivant : « Nom_souverain ({Période}) : ». Il faut sécuriser la capture de la période de règne pour ne pas récupérer toutes les périodes du document mais seulement celles correspondants à la période de règne d'un souverain en ajoutant la contrainte "précédée de Souverain".
Exemple :
" *Type de 771-793/4 : Charlemagne (768-814)...*"

■

■ *Formalisation :*

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

Exemple :
" *..Charlemagne (768-814)...*"
Correspondance : Charlemagne
Group #1 : 768
Group #2 : 814

■

Nature de la monnaie

■ *Vulgarisation :*

La nature de la monnaie est toujours un élément de l'ensemble {Denier, Obole, Monnaie D'Or, Faux obole, Monnaies de type indéterminé} suivi du nom d'un souverain. Il suffit donc d'utiliser une expression composé de tous les mots de l'ensemble.
Exemple :
" *Obole de Charles le Chauve*"

■

■ *Formalisation :*

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

Exemple :
" *Obole de Charles le Chauve*"

Correspondance : Charles le Chauve
Group #1 : Obole

■

Légende

■ *Vulgarisation :*

La légende est toujours placée à la ligne qui suit la nature de la monnaie. Le droit est situé au début de cette ligne et commence par zéro ou un caractère +. Ensuite, vient une suite d'espaces. Enfin, le revers vient se placer après zéro ou un caractère +.
Exemple :
"Denier de Charlemagne .....
+ CARLO 45E CROIX SIMPLE"

■

■ *Formalisation :*

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?(.*)\n`

On obtient une expression comme suit :

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

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

■

Types monétaire

■ *Vulgarisation :*

Le type monétaire est la concaténation de "Type de" avec la période d'émission. Cette annotation est similaire à la période d'émission mais appartient à une dimension différente.
Exemple :
" Type de 771-793/4 : Charlemagne (768-814),..."

■
■ *Formalisation :*

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

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

■

Collections, Trésors, Trouvailles

■ *Vulgarisation :*

Les collections, trésors et trouvailles sont chacun des ensembles d'information à annoter séparément. Les *mots* Collections, Trésors et 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 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 :

Cinquième partie

Conclusion