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

L'évolution des volumes d'informations sur internet provoque une évolution du web vers une approche dans laquelle chaque donnée acquiert un sens afin de rendre possible une interprétation du contenu par des machines. Cette évolution constitue le web sémantique.
Sa principale motivation est la recherche d'information sémantique.

Dans ce cadre, nous sommes intervenus pour répondre à l'appel d'offre de la maîtrise d'ouvrage.
L'objectif correspond à la première étape dans un processus de recherche d'informations sémantiques et d'indexation : l'annotation sémantique d'un document texte spécifique.

En découle les problématiques suivantes :

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

La résolution de ces différentes problématiques nous a amenés à organiser notre réflexion :
Premièrement, nous définirons clairement les demandes et leurs contextes à travers le cahier des charges. Deuxièmement, nous présenterons l'état des connaissances actuelles sur le sujet. Troisièmement, nous détaillerons notre principe de résolution du projet. Nous finirons le bilan et le retour d'expérience de ce 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.

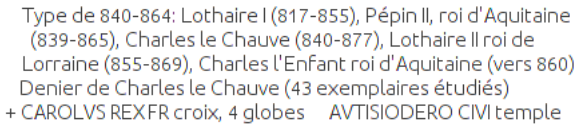
En résumé, les caractéristiques du projet sont :
- l'apprentissage et la compréhension du domaine considéré,
- l'étude des principes d'annotation de documents,
- le développement d'une chaîne d'annotation dans GATE,
- la mise en place d'une visualisation des résultats.

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

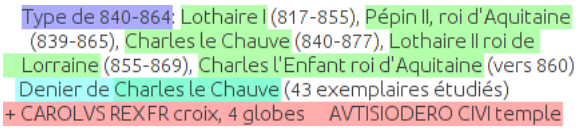
II Description

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



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 dû 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

Avant de démarrer le développement de la chaîne de traitement, nous devons nous familiariser avec notre environnement de travail. Celui-ci comprend l'outil GATE et le Numéraire carolingien. Ensuite, nous allons pouvoir définir nos annotations, leur domaine et choisir de quelle façon nous allons les capturer.

I Prise en main de l'environnement

1 GATE

2 Numéraire Carolingien

Au lancement du projet, la seule ressource a disposition été le Numéraire Carolingien au format papier. Il fallait le numériser et l'OCRiser.
Nous avons numérisé une cinquantaine de pages à la main pour les OCRiser automatiquement par la suite à l'aide de l'outil `tesseract`.

L'OCRisation s'est déroulée de la façon suivante :

La scanner que nous avons utilisé permettait d'obtenir une image pour chacune des pages au format PDF. Ensuite, étant donné que `tesseract` est plus performant et précis avec des fichiers TIFF, il a fallu convertir les fichiers PDF en TIFF. Cependant, quelques erreurs d'OCRisation sont apparues. Pour finir, `tesseract` nous donnait des fichiers TXT.
Les étapes de conversion du schéma précédent ont été réalisées à l'aide de script en langage SHELL :

1. (a)
```
for file in *.pdf
do
  convert -density 300 ../pdf/$file -depth 8 ‘basename $file .pdf‘.tiff
done
```

   (b) les pages impaires ont été numérisée à l'envers, il fallait les mettre dans le bon sens.

```
for file in *.tiff
do
    if [ $((‘basename $file .tiff‘ % 2 )) = 1 ]; then
        convert $file -rotate 180 $file;
    fi
```

```
    done
```

2. ```
   for file in img/*.tiff
   do
     tesseract $file txt/‘basename $file .tiff‘ -l fra
   done
   ```

II Définition des dimensions d'annotation et leur contenu

Dans l'introduction, nous avons vu qu'il y a trois dimensions d'annotation.

La dimension spatiale contient toutes les informations de lieux. Peu spécifique au domaine considéré.

La dimension temporelle contient toutes les informations de temps et de durées. Peu spécifique au domaine considéré.

La dimension thématique contient toutes les informations du domaine considéré.

III Première recherche d'entités nommées avec les gazetiers

Ateliers

■ *Vulgarisation :*

L'ouvrage est décomposé en ateliers qui donnent leur nom à chaque début de "partie" ou "chapitre". Ce nom correspond à un endroit de France ou pays limitrophes dans lequel est produite la monnaie. Ce lieu peut être une ville, un lieu-dit dont le nom peut ne plus exister. Il fût donc difficile de trouver un pattern via les expressions régulières. Nous avons alors dû construire ce gazetier à partir de la liste en début d'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 :*

Le nom des souverains ont des formats aussi divers 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
```

■

IV 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. En cas d'ambiguïté,Chacune d'elle peut 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 a la forme suivante : « Type de {Période} : » (Période étant l'annotation définie précédemment).
Il faut sécuriser la capture de la période d'émission en ajoutant la contrainte précédée de "Type de" afin de ne pas récupérer toutes les périodes du documents.
■

■ *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 a la forme suivante : « Nom_souverain ({Période}) : ». Il faut sécuriser la capture de la période de règne en ajoutant la contrainte précédée de "Souverain" afin de ne pas récupérer toutes les périodes du documents.
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 composée 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 sous la ligne de la nature de la monnaie. La légende du droit est située au début de cette ligne et commence par zéro ou un caractère +. Ensuite, vient une suite d'espaces. Enfin, la légende du 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 fois le symbole + et 2 espaces ou plus. C'est la légende du droit.

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

3. Capturer l'ensemble des caractères entre 0 ou 1 fois le 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 la dimension thématique.
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'informations 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 suivi 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