CONCEPTION ET DÉVELOPPEMENT D'UNE APPLICATION D'ANNOTATION THÉMATIQUE

dans l'environnement Gate

Charles Follet Roland Bary

Université de Pau et Pays de l'Adour

■ Extension du web vers un web dit de données: le web sémantique

- Extension du web vers un web dit de données: le web sémantique
- Recherche sémantique : activité du web sémantique

- Extension du web vers un web dit de données: le web sémantique
- Recherche sémantique : activité du web sémantique
- Annotation sémantique : étape du processus de recherche sémantique

- Extension du web vers un web dit de données: le web sémantique
- Recherche sémantique : activité du web sémantique
- Annotation sémantique : étape du processus de recherche sémantique
- Problématique : comment annoter sémantiquement un document texte non-structuré ?

- Extension du web vers un web dit de données: le web sémantique
- Recherche sémantique : activité du web sémantique
- Annotation sémantique : étape du processus de recherche sémantique
- Problématique : comment annoter sémantiquement un document texte non-structuré ?
- Quelles outils pour l'annotation?

- Extension du web vers un web dit de données: le web sémantique
- Recherche sémantique : activité du web sémantique
- Annotation sémantique : étape du processus de recherche sémantique
- Problématique : comment annoter sémantiquement un document texte non-structuré ?
- Quelles outils pour l'annotation?
- Quelle approche de conception?

SOMMAIRE

Le cahier des charges

Cadre d'analyse

Structure du numéraire

Annotations

Construction d'une règle JAPE

Visualisation

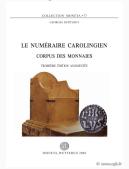
Contexte: Recherche d'informations sur le Numéraire Carolingien de Georges DEPEYROT.



Problème : répondre aux questions

- 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 ?

Contexte: Recherche d'informations sur le Numéraire Carolingien de Georges DEPEYROT.



Problème : répondre aux questions

- 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 ?

Solution: Utilisation d'un outil de Traitement automatique du langage.

Caractéristiques du projets:

■ apprentissage et compréhension du domaine considéré

Caractéristiques du projets:

- apprentissage et compréhension du domaine considéré
- étude des principes d'annotation de documents

Caractéristiques du projets:

- apprentissage et compréhension du domaine considéré
- étude des principes d'annotation de documents
- développement d'une chaîne d'annotation dans GATE

Caractéristiques du projets:

- apprentissage et compréhension du domaine considéré
- étude des principes d'annotation de documents
- développement d'une chaîne d'annotation dans GATE
- mise en place d'une interface web de visualisation des résultats

Scénario attendu: Texte en entrée de traitement

Auxerre (Yonne

Type de 840-864: Lothaire I (817-855), Pépin II, roi d'Aquitaine (839-865), Charles le Chauve (840-877), Lothaire II roi de Lorgier de 855-869), Charles l'Enfant roi d'Aquitaine (vers 860) Denier de Charles le Charve (43 exemplaires étudiés)

+ CAROLVS REX FR croix, 4 globes AVTISIODERO CM temple
Gariel XX 1, 2; Morrison-Grunthal 987-990; Depeyrot (I) (2) 96; Nombreuses variétés de légende, de
ponctuation et de cravure. Collections: Assen 1.54; Berlin: MEC 842 (1.59); New York 1,49; Prou

582 (1.63), 583 (1,46), Trésors: Groningen-1890 (vers 840), 1 ex. (MG 38: H. 39, daté de 845-846); Auberre (840-845), 1 ex. (D. 24); Tzummarum (840-845), 1 ex. (H. 22 daté de 840, ne mentionne so cette monnaie dans le catalogue); Cosne (840-845), 30 ex. (D. 109: H. 45, daté de 845-846, voir p. 93); Midlaren (855), 1 ex. (MG 55; H. 57 daté de 884); Nourry (869-875), 1 ex. (D. 42); H. 64 daté de 869-870); Roswinkel (875), 1 ex. (MG 66; H. 75 daté de 877) (1,55); Trouvailles: Auxerre, 1 ex. (H. 511)

Type de 864-875: Charles le Chauve (840-877), Louis le Germanique (840-876), Louis II le Bèque roi d'Aquitaine (867-

Denier de Charles le Chauve (22 exemplaires étudiés)
+ GRATIA DEL REX monogramme + AVTISIODER CIVITAS croix

877), de France (877-879)

Gariel XW 20-29; Morrison-Grunthal 991, 994, 996-1001; Depeyort (1) (2) 97; Nombreuses variétés de légende, de ponctuation et de gravuer (AWTSIODER CMI, Celletions Eerlin 1,77, 1.70, 1.59, 1.55; MEC 853 (1,78); Monnaie de Paris 105 (1,63); Prou S84 (1,58), 585 (1,59), 586 (1,79), 587 (1,72), 586 (1,77). Trésors: Fortiames/Mercurey (840-875), Lev. (D. 147 en mentionne pas cette monnaie; H. 66 daté de 875); Cosne (875-877), 2 ev. (D. 110; H. 67 daté de 875) (1,74, 1,73); Mortrieux-en-Sologne (887-889); A ev. (D. 231; H. 88 daté de 889); Gilsy (890), 3 ev. (D. 154; H. 90 daté de 888); Monchy-au-Bois (895-998); lev. (D. 217; H. 101 daté de 897-898); Montrieux-en-Sologne (907), Jev. (D. 229; H. 67 daté de 889).

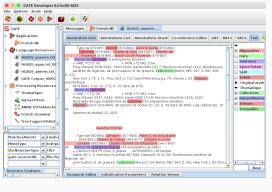
Denier de Charles le Chauve (1 exemplaire étudié) + GRATIA DEI REX monogramme + ALTIEI CIVITVS (rétro.) croix Gariel XXV 24: Morrison-Grunthal 991: Depeyor (1) (2) 98: Collections: Prou 589 (1.68).

Scénario attendu: Texte en sortie de traitement

Auxerre (Vonne) 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 Gariel XXI 1, 2: Morrison-Grunthal 987-990; Depeyrot (I) (2) 96; Nombreuses variétés de légende, de ponctuation et de gravure, Collections: Assen 1,54; Berlin: MEC 842 (1,59); New York 1,49; Prou 582 (1.63), 583 (1.46), Trésors: Groningen-1890 (vers 840), 1 ex. (MG 38; H, 39, daté de 845-846); Auxerre (840-845), 1 ex. (D. 24): Tzummarum (840-845), 1 ex. (H. 22 daté de 840, ne mentionne pas cette monnaie dans le catalogue): Cosne (840-845), 30 ex, (D. 109; H. 45, daté de 845-846, voir p. 93): Midlaren (855). 1 ex. (MG 55: H. 57 daté de 864): Nourray (869-875). 1 ex. (D. 243: H. 64 daté de 869-870); Roswinkel (875), 1 ex. (MG 66; H. 75 daté de 877) (1.55); Trouvailles: Auxerre, 1 ex. (H. 511) Type de 864-875: Charles le Chauve (840-877), Louis le Germanique (840-876), Louis II le Bèque roi d'Aquitaine (867-877), de France (877-879) Denier de Charles le Chauve (22 exemplaires étudiés) + GRATIA DEI REX monogramme + AVTISIODER CIVITAS croix Gariel XXV 20-23: Morrison-Grunthal 991, 994, 996-1001: Depeyrot (1) (2) 97: Nombreuses variétés de légende, de ponctuation et de gravure (AVTISIODER CIVI). 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: Fontaines/Mercurey (840-875), 1 ex. (D. 147 ne mentionne pas cette monnaie; H. 66 daté de 875); Cosne (875-877), 2 ex. (D. 110; H. 67 daté de 875) (1,74, 1,73); Montrieux-en-Sologne (887-898), 4 ex. (D. 231: H. 88 daté de 888); Glisv (890), 3 ex. (D. 154: H. 90 daté de 888). Monchy-au-Bois (895-898). 1 ex. (D. 217: H. 101 daté de 897-898): Montrieux-en-Sologne (900), 1 ex. (D. 229; H. 87 daté de 888). 98 Denier de Charles le Chauve (1 exemplaire étudié) + GRATIA DEI REX monogramme + ALTIFI CIVITVS (rétro.) croix Gariel XXV 24; Morrison-Grunthal 991; Depeyrot (1) (2) 98; Collections: Prou 589 (1,68).

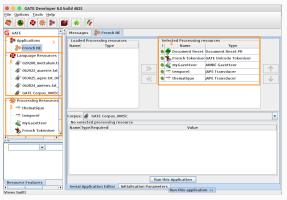
CADRE D'ANALYSE

CADRE D'ANALYSE - L'OUTIL GATE



- Outil de traitement automatique du langage (en JAVA).
- Université de Sheffield 1995.
- Principe: Chaîne de traitement (Pipeline).
- Plusieurs modules dans un pipeline.
- ANNIE: Pipeline par défaut pouvant servir de chaîne de départ

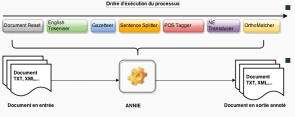
CADRE D'ANALYSE - L'OUTIL GATE



- 1- Chaîne de traitement
- 2- Textes et corpus à traiter
- 3- Modules d'annotation
- 4- Composition d'une chaîne de traitement

CADRE D'ANALYSE - L'OUTIL GATE

■ Exemple de la chaîne de traitement ANNIE



- Module Gazetter:
 Identification des noms d'entité dans le texte, sur la base d'un gazetier
 - Module NE Transducer: Repérage d'entités nommées inconnues en fonction de modèles d'extraction écrits en langage JAPE

CADRE D'ANALYSE - GAZETIERS

Bérenger I:valeur=Bérenger I:periode=888-924
Bérenger II:valeur=Bérenger II:periode=950-961
Carloman:valeur=Carloman:periode=768-771
Carloman:valeur=Carloman:periode=879-884
Charlemagne:valeur=Charlemagne:periode=768-814

- liste de noms d'entités (Villes, métiers,etc)
- permettant un repérage d'occurrences dans un texte
- possibilité d'associer des caractéristiques a chaque nom d'entité (séparateurs)
- Dans Gate: MajorType, MinorType

CADRE D'ANALYSE - LE FORMALISME JAPE

```
Phase: pourcent
Input: Lookup Token
Options: control = appelt
/******* MACROS */
Macro: AMOUNT NUMBER
(({Token.kind == number}
({Token.string == ","} | {Token.string == "."})
{Token.kind == number}
)?
({Token.string == "-"}{Token.kind == number})?))
Macro: PERCENT(
{Token.string == "%"} |
{Token.string == "pourcent"}|
({Token.string == "pour"}{Token.string == "cent"})14)
Macro: LIAISON PERCENT DE ET(
{Token.string == "de"}|
{Token.string == "et"}|
{Token.string == ","})
/***** Regle */
Rule: Percent_number
(((AMOUNT NUMBER)
({Token.string == "a"})
)?(AMOUNT_NUMBER)(PERCENT))
:pourcent
-->:pourcent.Percent =
{Kind="pourcent".Rule="Percent number"}
```

- Règles JAPE en deux blocs
- Partie gauche (LHS): définition d'un motif d'annotation à repérer
- Partie droite (RHS): Opérations sur un motif (Code JAVA éventuel)
- Utilisation d'un système de Macros



EXEMPLE

Agen (Lot-et-Garonne)

L'ATELIER

Agen (Lot-et-Garonne)

16

LE TYPE

Agen (Lot-et-Garonne)

Type de 793/4-812: Charlemagne (768-814)

Denier de Charlemagne (24 exemplaires étudiés)

Gariel XXII 26; Morrison-Grunthal 177-179; Depeyrot (1) (2) 1; Collections: Berlin 1,64, 1,45; Bruxelles 1,25, 1,25; Charleville-Mézières 1,56; Copenhague 1,31; Garett 1,73; Grenoble; MEC 735 (1,57), 736 (1,24); Monnaie de Paris 88 (1,55); New York 1,64, 1,57; Prou 792 (1,36), 793 (1,60), 794 (1,50) Trésors: Biebrich (790-814), 2 ex. (MG 11, Vôlkers, p. 182) (1,60); Ibersheim (814), 1 ex. (MG 13, Vôlkers, p. 186) (1,64); Dorestad (822), 4 ex. (MG 18, Vôlkers, p. 139; H. 7) (1,35);

LA NATURE

Agen (Lot-et-Garonne)

Type de 793/4-812: Charlemagne (768-814)

Denier de Charlemagne (24 exemplaires étudiés)

+ CARLVS REX FR croix + AGINNO monogramme

Gariel XXII 26; Morrison-Grunthal 177-179; Depeyrot (1) (2) 1; Collections: Berlin 1,64, 1,45; Bruxelles 1,25, 1,25; Charleville-Mézières 1,56; Copenhague 1,31; Garett 1,73; Grenoble; MEC 735 (1,57), 736 (1,24); Monnaie de Paris 88 (1,55); New York 1,64, 1,57; Prou 792 (1,36), 793 (1,60), 794 (1,50) Trésors: Biebrich (790-814), 2 ex. (MG 11, Vôlkers, p. 182) (1,60); Ibersheim (814), 1 ex. (MG 13, Vôlkers, p. 186) (1,64); Dorestad (822), 4 ex. (MG 18, Vôlkers, p. 139; H. 7) (1,35);

LA LÉGENDE

Agen (Lot-et-Garonne)

Type de 793/4-812: Charlemagne (768-814) Denier de Charlemagne (24 exemplaires étudiés)

+ CARLVS REX FR croix + AGINNO monogramme

Gariel XXII 26; Morrison-Grunthal 177-179; Depeyrot (1) (2) 1; Collections: Berlin 1,64, 1,45; Bruxelles 1,25, 1,25; Charleville-Mézières 1,56; Copenhague 1,31; Garett 1,73; Grenoble; MEC 735 (1,57), 736 (1,24); Monnaie de Paris 88 (1,55); New York 1,64, 1,57; Prou 792 (1,36), 793 (1,60), 794 (1,50) Trésors: Biebrich (790-814), 2 ex. (MG 11, Vôlkers, p. 182) (1,60); Ibersheim (814), 1 ex. (MG 13, Vôlkers, p. 186) (1,64); Dorestad (822), 4 ex. (MG 18, Vôlkers, p. 139; H. 7) (1,35);

LES COLLECTIONS

Agen (Lot-et-Garonne)

Type de 793/4-812: Charlemagne (768-814)

Denier de Charlemagne (24 exemplaires étudié:

Gariel XXII 26; Morrison-Grunthal 177-179; Depeyrot (1) (2) 1; Collections: Berlin 1,64, 1,45; Bruxelles 1,25, 1,25; Charleville-Mézières 1,56; Copenhague 1,31; Garett 1,73; Grenoble; MEC 735 (1,57), 736 (1,24); Monnaie de Paris 88 (1,55); New York 1,64, 1,57; Prou 792 (1,36), 793 (1,60), 794 (1,50) Trésors: Biebrich (790-814), 2 ex. (MG 11, Völkers, p. 182) (1,60); Ibersheim (814), 1 ex.

LES TRÉSORS

Agen (Lot-et-Garonne)

21

LES TROUVAILLES

Agen (Lot-et-Garonne)

RÉSUMÉ

```
Agen (Lot-et-Garonne) ATELIER

Type de 793/4-812: Charlemagne (768-814) TYPE

Denier de Charlemagne (24 exemplaires étudiés) NATURE

+ CARLVS REX FR croix + AGINNO monogramme Légende

Gariel XXII 26; Morrison-Grunthal 177-179; Depeyrot (1) (2) 1; Collections: Berlin 1,64, 1,45;

Bruxelles 1,25, 1,25; Charleville-Mézières 1,56; Copenhague 1,31; Garett 1,73; Grenoble; MEC 735

(1,57), 736 (1,24); Monnaie de Paris 88 (1,55); New York 1,64, 1,57; Prou 792 (1,36), 793 (1,60), 794 (1,50) COLLECTIONS Trésors: Biebrich (790-814), 2 ex. (MG 11, Vôlkers, p. 182) (1,60); Ibersheim (814), 1 ex.

(MG 13, Vôlkers, p. 186) (1,64); Dorestad (822), 4 ex. (MG 18, Vôlkers, p. 139; H. 7) (1,35), TRÉSORS

Trouvailles: Bolsward, 1 ex. (H. 528) (1,26) TROUVAILLES.
```

ANNOTATIONS

DEUX FAÇONS D'ANNOTER

GAZETIER liste de valeurs.

- · Type de 754/5-768
- · Type de 768-771
- · Type de 771-793/4
- · Type de 793/4-812
- · Type de 812-814

٠.

RÈGLE JAPE

règle basée sur une expression régulière.

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

// Exemple : Type de 758/9-762
Rule: PeriodeEmissionRule
(
    CHAINE_DEBUT ({Periode}):p
):PeriodeEmission
-->:PeriodeEmission.PeriodeEmission = {
    Kind = "PeriodeEmission",
    D1 = :p.Periode.D1, D2 = :p.Periode.D2
}
```

DEUX FAÇONS D'ANNOTER

GAZETIER liste de valeurs.

- · Type de 754/5-768
- · Type de 768-771
- · Type de 771-793/4
- · Type de 793/4-812
- · Type de 812-814
- ---

RÈGLE JAPE

règle basée sur une expression régulière.

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

// Exemple : Type de 758/9-762
Rule: PeriodeEmissionRule
(
    CHAINE_DEBUT ({Periode}):p
):PeriodeEmission
->:PeriodeEmission.PeriodeEmission = {
    Kind = "PeriodeEmission",
    D1 = :p.Periode.D1, D2 = :p.Periode.D2
}
```

CONSTRUCTION D'UNE RÈGLE JAPE

1. Langage naturel

- 1. Langage naturel
- 2. Langage régulier

- 1. Langage naturel
- 2. Langage régulier
- 3. Langage de programmation

- 1. Langage naturel
- 2. Langage régulier
- 3. Langage de programmation

français

- 1. Langage naturel
- 2. Langage régulier
- 3. Langage de programmation

français expressions régulières

- 1. Langage naturel
- 2. Langage régulier
- 3. Langage de programmation

français

expressions régulières

JAPE

Annotation : La période d'émission

Exemple : *Type de 793/4-812*

Type de 793/4-812

1. LANGAGE NATUREL

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

Il faut assurer la capture de la période d'émission en ajoutant la contrainte précédée de "Type de" afin de ne pas récupérer toutes les périodes du document.

Type de 793/4-812

1. LANGAGE NATUREL

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

Il faut assurer la capture de la période d'émission en ajoutant la contrainte précédée de "Type de" afin de ne pas récupérer toutes les périodes du document.

Type de **793/4**-812

2. LANGAGE RÉGULIER

Type de **793/4-**812

2. LANGAGE RÉGULIER

Type de 793/4-812

2. Langage régulier

Type de 793/4-812

2. LANGAGE RÉGULIER

```
Rule: PeriodeRule
((DATE):d1({Token.string =="-"})(DATE):d2)
:Periode -->:Periode{/*...*/}

Rule: PeriodeEmissionRule
(CHAINE_DEBUT ({Periode}):p)
:PeriodeEmission
-->:PeriodeEmission.PeriodeEmission = {/*..*/}
```

```
Rule: PeriodeRule
((OATE):d1({Token.string =="-"})(DATE):d2)
:Periode -->:Periode{/*...*/}

Rule: PeriodeEmissionRule
(CHAINE_DEBUT ({Periode}):p)
:PeriodeEmission
-->:PeriodeEmission.PeriodeEmission = {/*...*/}
```

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

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)
Macro:DATE_IMPRECISE
(TROIS_NOMBRES)
Macro:DATE
(DATE_PRECISE | DATE_IMPRECISE)
```

```
Rule: PeriodeRule
((DATE):d1({Token.string =="-"})(DATE):d2)
:Periode -->:Periode{/*...*/}

Rule: PeriodeEmissionRule
(CHAINE_DEBUT ({Periode}):p)
:PeriodeEmission
-->:PeriodeEmission.PeriodeEmission = {/*..*/}
```

Résultat d'annotation

<PeriodeEmission>Type de 793/4-812/PeriodeEmission>

Résultat d'annotation

<PeriodeEmission>Type de 793/4-812/PeriodeEmission>

Comment le visualiser de façon élégante ?

Résultat d'annotation

<PeriodeEmission>Type de 793/4-812</PeriodeEmission>

Comment le visualiser de façon élégante ?

XSLT



CONCLUSION

Retours d'expérience sur:

■ La gestion du temps

CONCLUSION

Retours d'expérience sur:

- La gestion du temps
- Respect des spécifications du cahier de charge

CONCLUSION

Retours d'expérience sur:

- La gestion du temps
- Respect des spécifications du cahier de charge
- Organisation du projet (gestion des réunions)

MERCI POUR VOTRE ATTENTION.

QUESTIONS ?