Méthodes d'étiquetage sémantique compatibles avec les données ouvertes

Applications de la plateforme Wikimeta.

Eric Charton

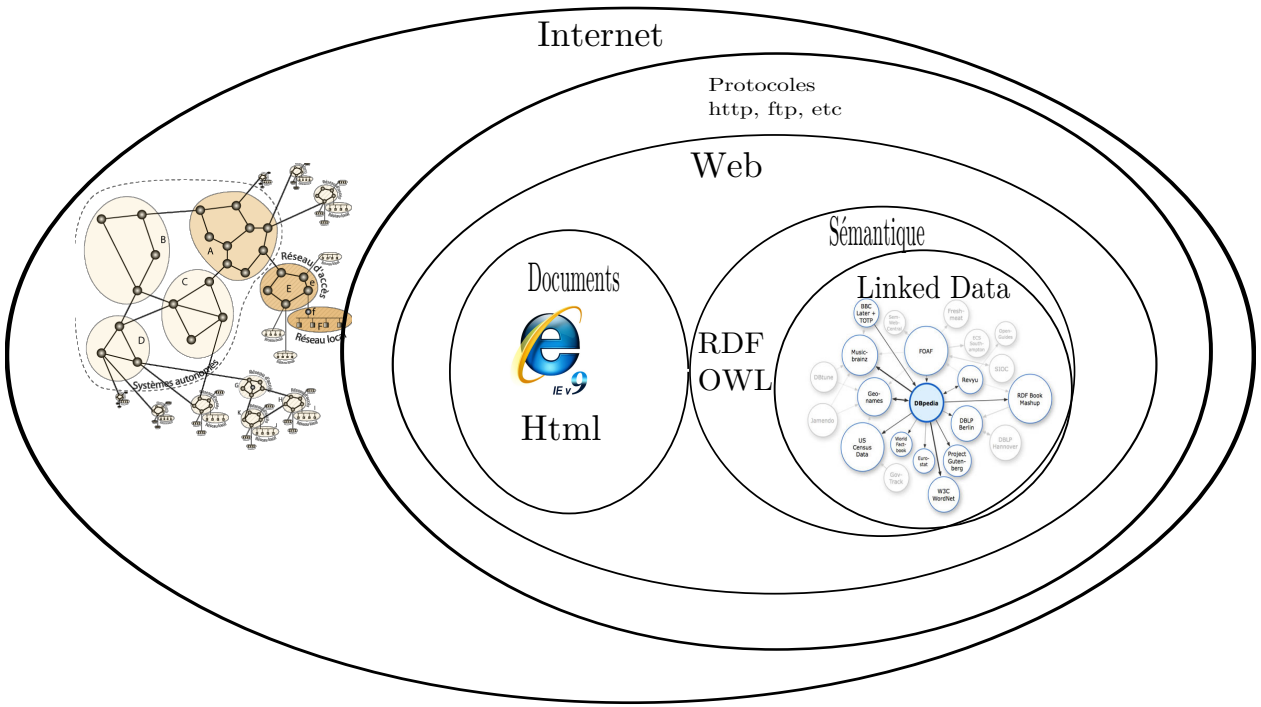
ÉCOLE POLYTECHNIQUE MONTRÉAL

Wikimeta

Plan

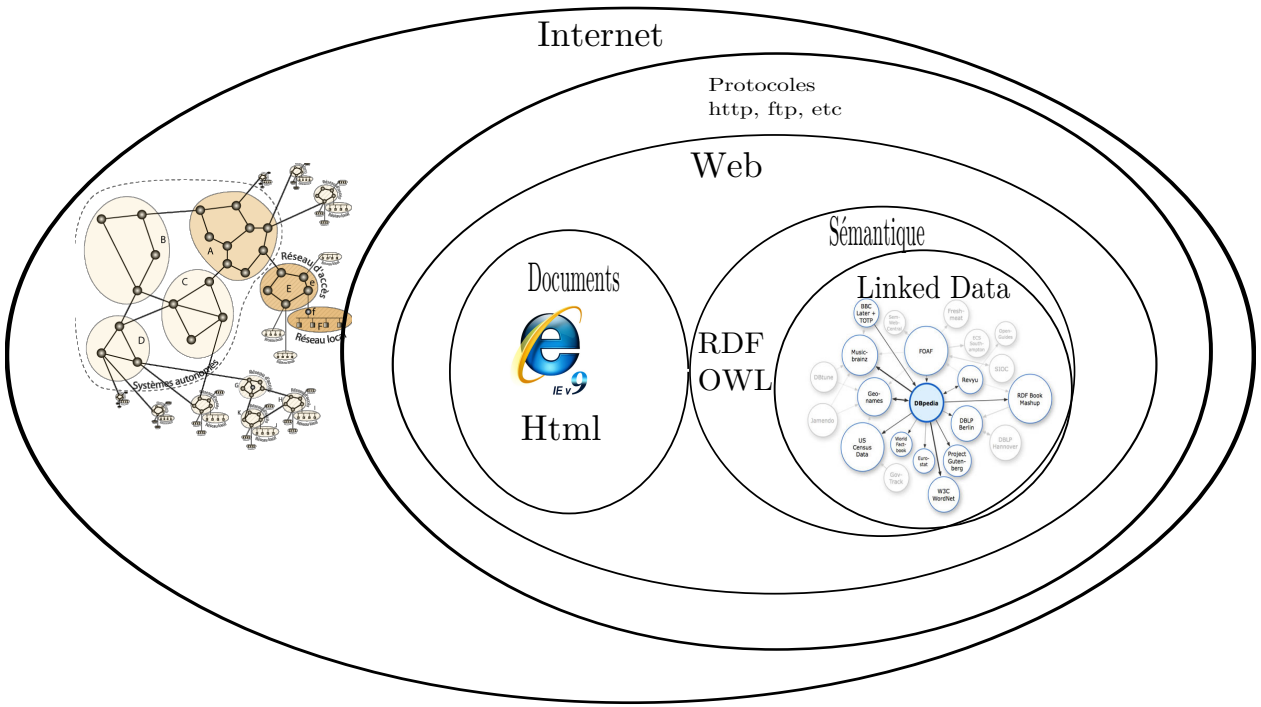
Qu'est ce que le web sémantique ?

Un ensemble de standards et de méthodes qui régissent la publications de données sémantiques sur le web.



Qu'est ce que le web sémantique ?

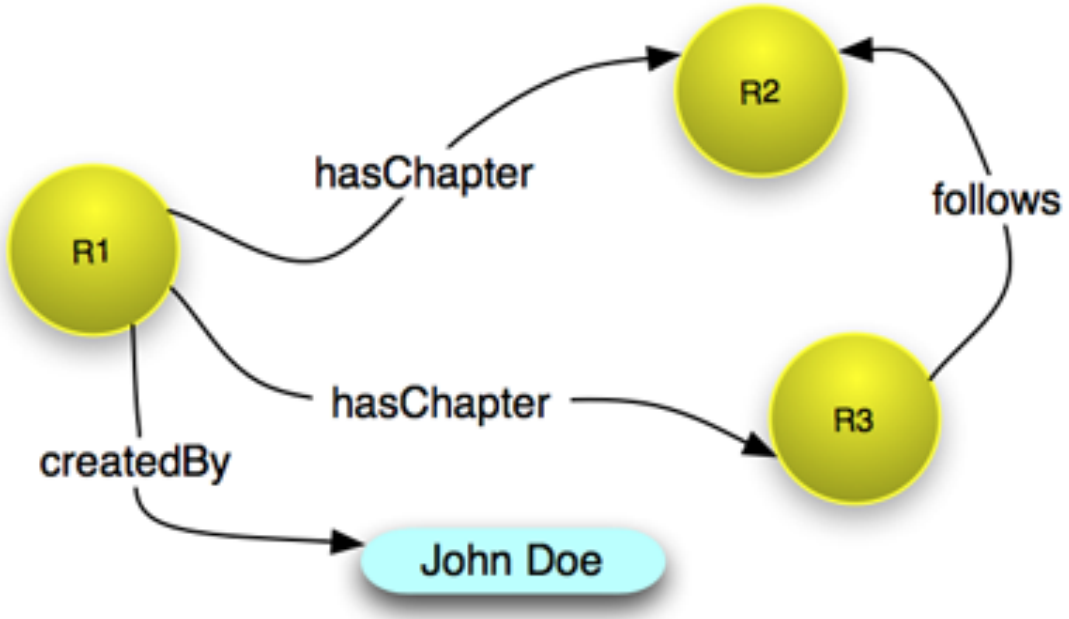
Un ensemble de standards et de méthodes qui régissent la publications de données sémantiques sur le web.



Le réseau Linked Data est un sous espace du web semantique qui décrit virtuellement toute connaissance humaine.

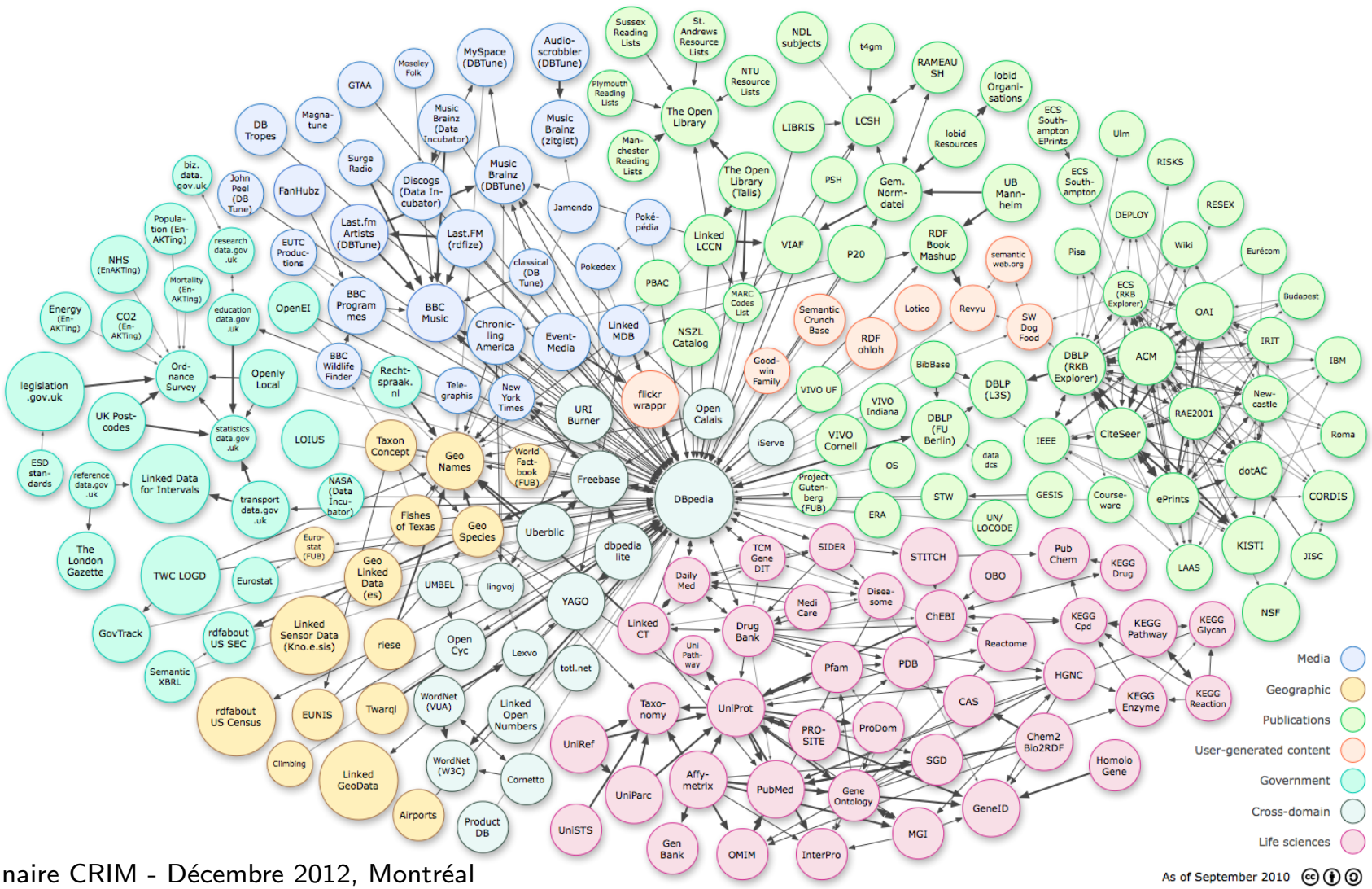
Structure du web semantique

Les triplets RDF sont utilisés pour décrire des concepts et leur propriétés ...



Subject	Predicate	Object
R1	hasChapter	R2
R1	hasChapter	R3
R3	follows	R2
R1	createdBy	"John Doe"

Le réseau Linked Data (en September 2010)



La relation aux données ouvertes

Les données ouvertes sont des informations structurées, librement diffusées.

Exemples de données ouvertes

- Données économiques (ONU,Gouvernements)
- Informations administrative (Statistiques)
- Informations géopolitiques (UNESCO Pisa)

Les données ouvertes ne sont pas obligatoirement standardisées.

La relation aux données ouvertes

Les données ouvertes sont des informations structurées, librement diffusées.

Exemples de données ouvertes

- Données économiques (ONU,Gouvernements)
- Informations administrative (Statistiques)
- Informations géopolitiques (UNESCO Pisa)

Les données ouvertes ne sont pas obligatoirement standardisées.

Exemples de formats de diffusion données ouvertes

- CSV, DOC, SQL, **RDF**

Les données ouvertes du web sémantique

Les données ouvertes du web sémantique

Exemples de données et de leur rapport au Web Sémantique

- Wikipedia ← DBPedia
- CIA World Fact Book ← CIA World Factbook en RDF
- DBLP ← DBLP RDF

Des enjeux importants!

Plus de 20 milliards de triplets RDF en 2010. Aujourd'hui, la taille du Web Semantique est devenue un sujet de recherche !

- Le mouvement de l'Open Data touche désormais la plupart des administrations.
- La plupart des données ouvertes devraient migrer à terme vers les formats du Web Semantique.
- De nouvelles applications doivent être élaborées.

La mise en relation des documents avec le Web Sémantique par l'étiquetage est l'une des briques de base des futures applications.

Des applications à inventer

Relier des connaissances à des contenus. Analyser automatiquement des documents.

Des applications à inventer

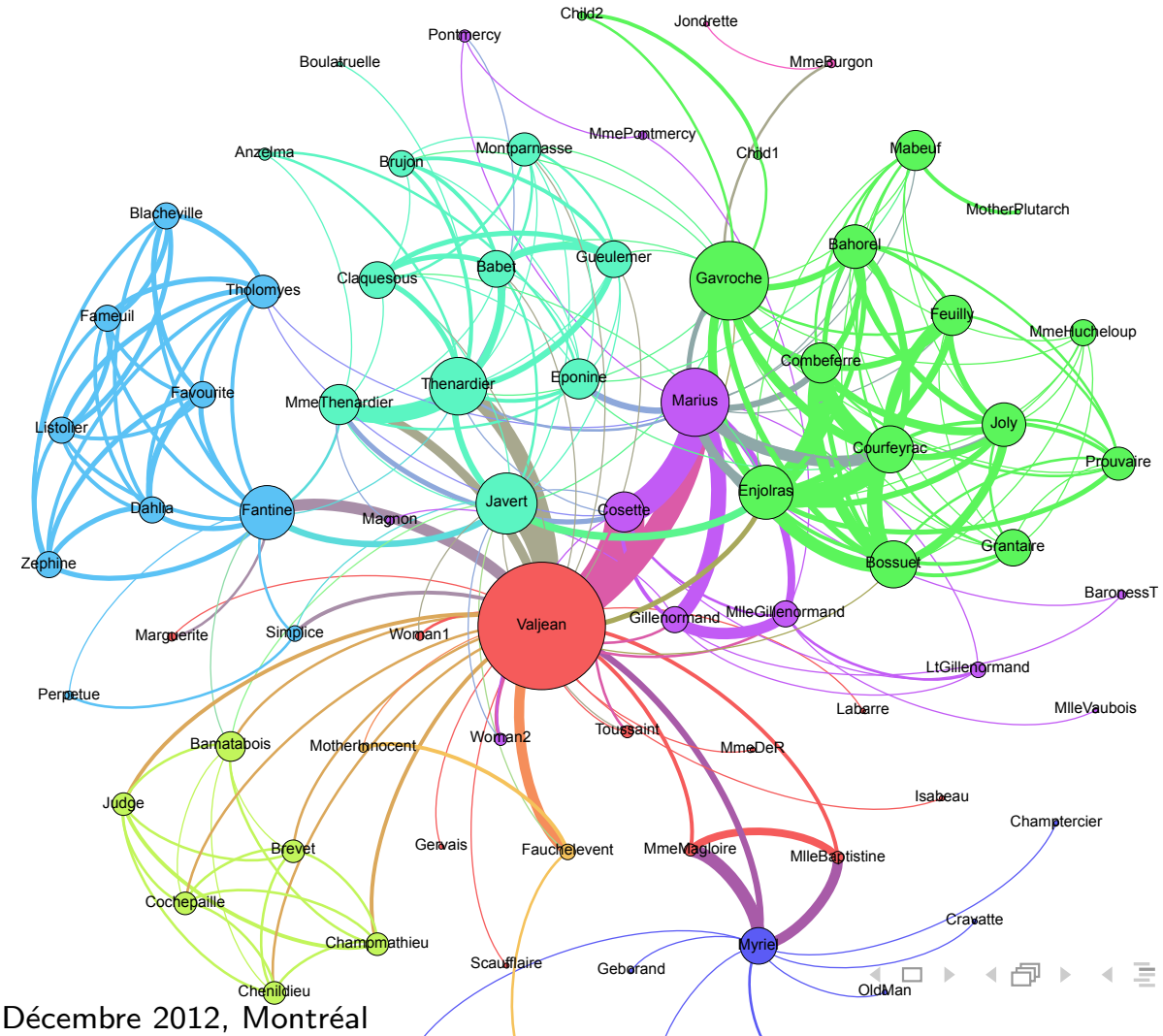
Relier des connaissances à des contenus. Analyser automatiquement des documents.

Applications

- Traiter et vérifier des flux d'informations
- Résumer des contenus
- Journalisme de données
- Aide à la décision
- Enrichissement des contenus
- Veille technologique
- Ingéniérie d'affaire

Exemples applicatifs

Exemple issu des Misérables de Victor Hugo, applicable aux mails, aux corpus...



Exemples applicatifs

Analyse d'une nouvelle

Wikimeta

Account and API

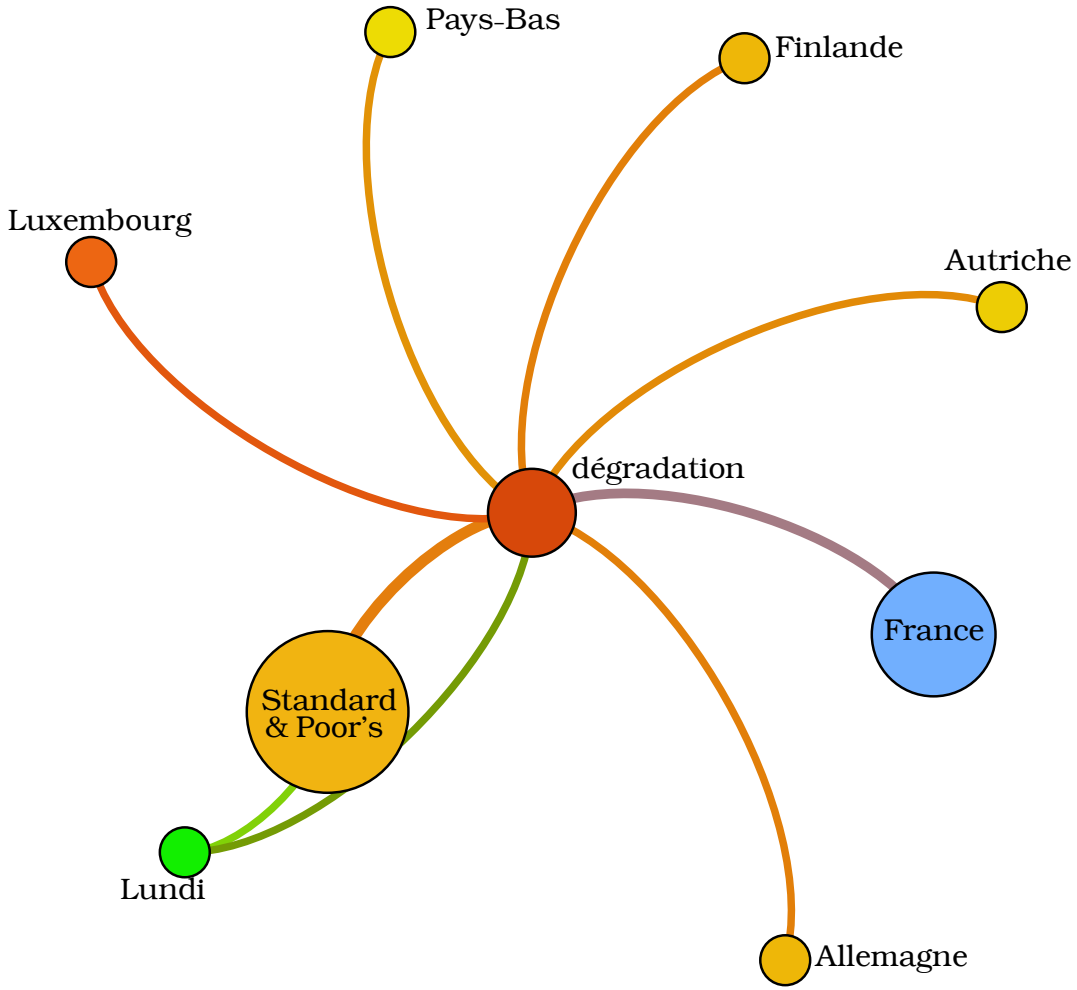
Semantic Tagger

Standard & Poor [ORG]'s [PROD] a lancé lundi [TIME] une menace de dégradation du triple A à la France [LOC], qui n'est pas la seule visée. L'agence de notation est sur le point de « mettre sous surveillance avec une implication négative » pas moins de six pays dont l'Autriche [LOC], la Finlande [LOC], le Luxembourg [ORG], les Pays-Bas [LOC] mais aussi l'Allemagne [LOC]. SUR LE MÊME SUJET Accord [ORG] Sarkozy-Merkel: l'opposition dénonce « perte de souveraineté » et « austérité » Hollande [PERS] attaque la politique de Sarkozy [PERS], un « échec » Standard & Poor [ORG] 's menace le triple A français et allemandEuro: Merkel [PERS] et Sarkozy [PERS] veulent imposer leur compromis

Les Etats [LOC] auraient été prévenus qu'ils risquaient de perdre leur AAA dans les quatre-vingt [TIME]dix jours [AMOUNT] et basculer dans la catégorie AA [ORG] +. Pour la France [LOC], S & P [ORG] envisage même une dégradation plus sévère de deux crans d'un coup.

Exemples applicatifs

Transformation pour analyse.



Le problème de l'annotation sémantique.

Analyse sémantique et annotation sémantique

Je veux réserver une chambre dans l'hôtel Beau-Soleil le mois prochain

Analyse sémantique et annotation sémantique

Je veux réserver une chambre dans l'hôtel Beau-Soleil le mois prochain

Analyse sémantique

Assigner un sens aux relations entre mots contenus dans la phrase (circonstances, acteurs, évènements)

Action $\rightarrow$ réserver; délais$\rightarrow$ 30 jours

Analyse sémantique et annotation sémantique

Je veux réserver une chambre dans l'hôtel Beau-Soleil le mois prochain

Analyse sémantique

Assigner un sens aux relations entre mots contenus dans la phrase (circonstances, acteurs, évènements)

Action → réserver; délais→ 30 jours

Annotation sémantique

Assigner une identité aux objets textuels contenus dans la phrase et obtenir leur propriétés

EN → Hôtel Beau Soleil; Propriétés → particularités des chambres, prix, disponibilités ...

Analyse sémantique et annotation sémantique: deux natures différentes.

Analyse sémantique et annotation sémantique: deux natures différentes.

Analyse sémantique

Obtenir de l'information sur le texte analysé et ses objectifs.

Analyse sémantique et annotation sémantique: deux natures différentes.

Analyse sémantique

Obtenir de l'information sur le texte analysé et ses objectifs.

Annotation sémantique

Introduire une **information extérieure au texte**.

Analyse sémantique et annotation sémantique: deux natures différentes.

Analyse sémantique

Obtenir de l'information sur le texte analysé et ses objectifs.

Annotation sémantique

Introduire une **information extérieure au texte**.

→ Principe de base de l'annotation sémantique

Détecter une entité → lui associer une identité unique → collecter des données extérieures.

Les différentes natures d'un objet textuel

L'Entité Nommée (EN) est le premier niveau de l'annotation sémantique.

Nature et limitations particulière de l'entité nommée

- La tâche de détection d'EN consiste à attribuer une classe à une séquence de mots: pers.hum,loc.fac, org.com.
- Le label de classe étant unique, il est impossible de l'utiliser pour définir les attributs sémantiques de l'entité.

Exemple: *Montréal* → loc → loc.admi → loc.admi.**?** population ? fondateurs ? et quel Montréal ?

Exemple: identification d'une entité candidate

La position de Paris à un carrefour entre les itinéraires commerciaux terrestres et fluviaux au cœur d'une riche région agricole en a fait une des principales villes de France au cours du Xe siècle, avec des palais royaux, de riches abbayes et une cathédrale

Détection des entités nommées

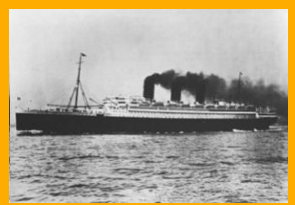
Exemple: désambiguïsation de classe

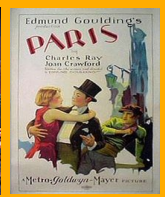
La position de Paris à un carrefour entre les itinéraires commerciaux terrestres et fluviaux au cœur d'une riche région agricole en a fait une des principales villes de France au cours du Xe siècle, avec des palais royaux, de riches abbayes et une cathédrale

?

Détection de classe d'une entité nommée





Un astéroïde, une ville, un navire, un produit ...

Exemple: choix de la classe finale

La position de Paris à un carrefour entre les itinéraires commerciaux terrestres et fluviaux au cœur d'une riche région agricole en a fait une des principales villes de France au cours du Xe siècle, avec des palais royaux, de riches abbayes et une cathédrale

`Une ville -> LOC.ADMI`

Exemple: une ambiguïté subsiste...

La position de Paris à un carrefour entre les itinéraires commerciaux terrestres et fluviaux au cœur d'une riche région agricole en a fait une des principales villes de France au cours du Xe siècle, avec des palais royaux, de riches abbayes et une cathédrale

`Une ville -> LOC.ADMI`

Mais quelle ville ?

... à l'intérieur de la classe.

La position de Paris à un carrefour entre les itinéraires commerciaux terrestres et fluviaux au cœur d'une riche région agricole en a fait une des principales villes de France au cours du Xe siècle, avec des palais royaux, de riches abbayes et une cathédrale


Paris (Ontario)


Paris


Paris (Tenessee)


Paris (Kentucky)


Paris (Maine)

Paris (Idaho)

L'ambiguïté est la limitation principale des entités nommées pour les traitements évolués.

Exemple de problème difficile

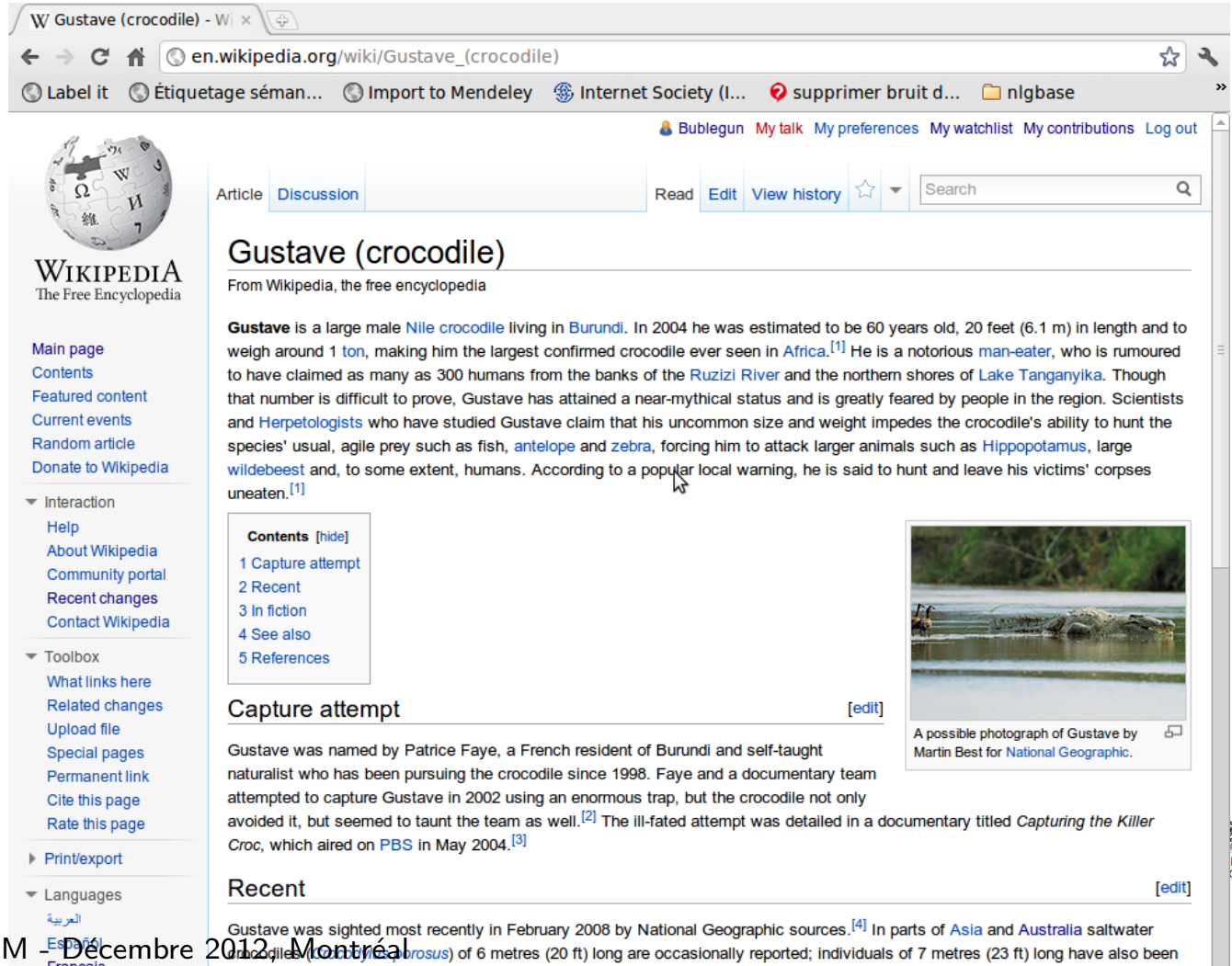
Gustave le crocodile

C'est au Burundi [LOC], sur les rives du lac Tanganyika [LOC] et à proximité, qu'un crocodile géant a semé la terreur pendant 10 [AMOUNT]ans [TIME]. Ce crocodile du Nil [LOC], surnommé Gustave [PERS] par les scientifiques, aurait tué environ 300 personnes [AMOUNT]. Un documentaire diffusé sur France 3 [ORG] nous a permis de découvrir ce monstre qui n'a pu d'ailleurs être capturé et est probablement mort aujourd'hui [TIME]. Cette émission bien qu'intéressante ne nous délivre quasiment aucune information scientifique sur ce crocodile. S'agit-il d'un crocodile du Nil [LOC] hors norme, de la même manière que certaines personnes ont des mensurations supérieures au standard? Possède t-il des caractéristiques particulières qui en feraient un crocodile d'une espèce non répertoriée?

En comparant les images tournées au Burundi [LOC] et celles de crocodiles du Nil [LOC], nous pourrons peut-être répondre à ces questions.

Exemple de problème difficile

Gustave le crocodile



Wikimeta: une architecture pour l'étiquetage sémantique

Deux problématiques de reconnaissances

Reconnaissance d'entités nommées

Deux problématiques de reconnaissances

Reconnaissance d'entités nommées

- Un classifieur (CRF, SVM) localise et classe les EN d'après leur contexte

Deux problématiques de reconnaissances

Reconnaissance d'entités nommées

- Un classifieur (CRF, SVM) localise et classe les EN d'après leur contexte
- Un système à base de lexique de détection appris automatiquement complète le classifieur.

Deux problématiques de reconnaissances

Reconnaissance d'entités nommées

- Un classifieur (CRF, SVM) localise et classe les EN d'après leur contexte
- Un système à base de lexique de détection appris automatiquement complète le classifieur.
- Le système est capable d'étiqueter une quantité limitée de classes (4/250) avec un contenu générique.

Deux problématiques de reconnaissances

Reconnaissance d'entités nommées

- Un classifieur (CRF, SVM) localise et classe les EN d'après leur contexte
- Un système à base de lexique de détection appris automatiquement complète le classifieur.
- Le système est capable d'étiqueter une quantité limitée de classes (4/250) avec un contenu générique.

(ex Une entreprise ORG.COM) $\rightarrow$ mots de contexte génériques $\rightarrow$ *Nyse, CAC40, Bilan, Effectifs, Shares, Revenu* etc)..

Deux problématiques de reconnaissances

Reconnaissance d'entités nommées

- Un classifieur (CRF, SVM) localise et classe les EN d'après leur contexte
- Un système à base de lexique de détection appris automatiquement complète le classifieur.
- Le système est capable d'étiqueter une quantité limitée de classes (4/250) avec un contenu générique.

(ex Une entreprise ORG.COM) $\rightarrow$ mots de contexte génériques $\rightarrow$ *Nyse, CAC40, Bilan, Effectifs, Shares, Revenu* etc)..

Étiquetage smantique

Deux problématiques de reconnaissances

Reconnaissance d'entités nommées

- Un classifieur (CRF, SVM) localise et classe les EN d'après leur contexte
- Un système à base de lexique de détection appris automatiquement complète le classifieur.
- Le système est capable d'étiqueter une quantité limitée de classes (4/250) avec un contenu générique.

(ex Une entreprise ORG.COM) $\rightarrow$ mots de contexte génériques $\rightarrow$ *Nyse, CAC40, Bilan, Effectifs, Shares, Revenu* etc)..

Étiquetage smantique

- Pour chaque entité il existe un graphe sémantique sur le réseau LinkedData.

Deux problématiques de reconnaissances

Reconnaissance d'entités nommées

- Un classifieur (CRF, SVM) localise et classe les EN d'après leur contexte
- Un système à base de lexique de détection appris automatiquement complète le classifieur.
- Le système est capable d'étiqueter une quantité limitée de classes (4/250) avec un contenu générique.

(ex Une entreprise ORG.COM) $\rightarrow$ mots de contexte génériques $\rightarrow$ *Nyse, CAC40, Bilan, Effectifs, Shares, Revenu* etc)..

Étiquetage smantique

- Pour chaque entité il existe un graphe sémantique sur le réseau LinkedData.
- Le contexte est spécifique: à chaque identité d'une NE correspond un contexte unique.

Deux problématiques de reconnaissances

Reconnaissance d'entités nommées

- Un classifieur (CRF, SVM) localise et classe les EN d'après leur contexte
- Un système à base de lexique de détection appris automatiquement complète le classifieur.
- Le système est capable d'étiqueter une quantité limitée de classes (4/250) avec un contenu générique.

(ex Une entreprise ORG.COM) → mots de contexte génériques → *Nyse, CAC40, Bilan, Effectifs, Shares, Revenu* etc)..

Étiquetage smantique

- Pour chaque entité il existe un graphe sémantique sur le réseau LinkedData.
- Le contexte est spécifique: à chaque identité d'une NE correspond un contexte unique.

(ex Paris (France) → mots contextuels personnalisés → *Seine, Tour Eiffel* etc).

Proposition

Séparer les deux tâches de reconnaissance d'entités et d'étiquetage sémantique. Le système cherche à associer à une NE détectée un lien sémantique.

Proposition

Séparer les deux tâches de reconnaissance d'entités et d'étiquetage sémantique. Le système cherche à associer à une NE détectée un lien sémantique.

Linked Data Interface (LDI)

Proposition

Séparer les deux tâches de reconnaissance d'entités et d'étiquetage sémantique. Le système cherche à associer à une NE détectée un lien sémantique.

Linked Data Interface (LDI)

- La **Linked Data Interface (LDI)** est une ressource statistique.

Proposition

Séparer les deux tâches de reconnaissance d'entités et d'étiquetage sémantique. Le système cherche à associer à une NE détectée un lien sémantique.

Linked Data Interface (LDI)

- La **Linked Data Interface (LDI)** est une ressource statistique.
- Elle contient pour chaque identité sémantique tous les mots de contextes possibles et leurs poids TF.IDF.

Proposition

Séparer les deux tâches de reconnaissance d'entités et d'étiquetage sémantique. Le système cherche à associer à une NE détectée un lien sémantique.

Linked Data Interface (LDI)

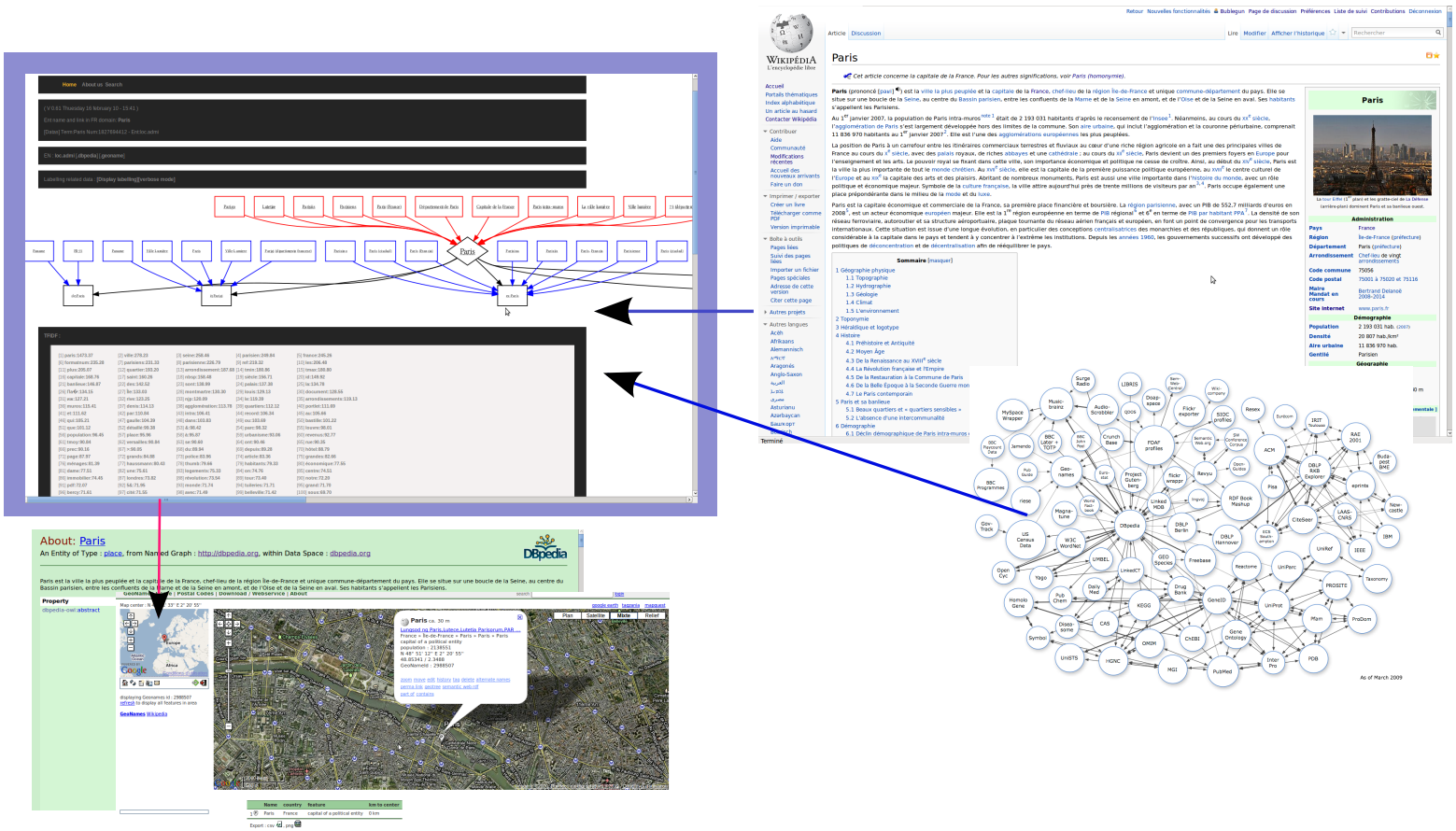
- La **Linked Data Interface (LDI)** est une ressource statistique.
- Elle contient pour chaque identité sémantique tous les mots de contextes possibles et leurs poids TF.IDF.
- Pour chaque entité sémantique, la LDI inclut un ou plusieurs lien vers le réseau LinkedData.

Linked Data Interface (LDI)

Construite d'après des ressources du web

- Wikipedia fournit 3.9M entités sémantiques avec leurs mots contextuels.
- Chaque entité sémantique est associée à un inventaire de formes de surfaces possibles (ex Lyon, Cité des Gaules, Ville de Lyon).
- Une table de correspondance entre Wikipédia et DBpedia fournit le lien entre l'entité sémantique et un point d'entrée du Web sémantique.

Construction de la Linked Data Interface



Pipeline du système

- L'étiqueteur d'entité nommées localise les EN dans le texte.

Pipeline du système

- L'étiqueteur d'entité nommées localise les EN dans le texte.
- La forme de surface de l'EN est utilisée pour localiser une ou plusieurs entités sémantiques de la LDI.

Pipeline du système

- L'étiqueteur d'entité nommées localise les EN dans le texte.
- La forme de surface de l'EN est utilisée pour localiser une ou plusieurs entités sémantiques de la LDI.
  - Une mesure de similarité est réalisée entre le contexte de l'EN et les sacs de mots de la LDI.

Pipeline du système

- L'étiqueteur d'entité nommées localise les EN dans le texte.
- La forme de surface de l'EN est utilisée pour localiser une ou plusieurs entités sémantiques de la LDI.
  - Une mesure de similarité est réalisée entre le contexte de l'EN et les sacs de mots de la LDI.
  - Si plus d'une candidate existe dans la LDI (ex *Paris (france)*, *Paris (Ontario)* ...), le meilleure score de similarité indique la meilleure ES.

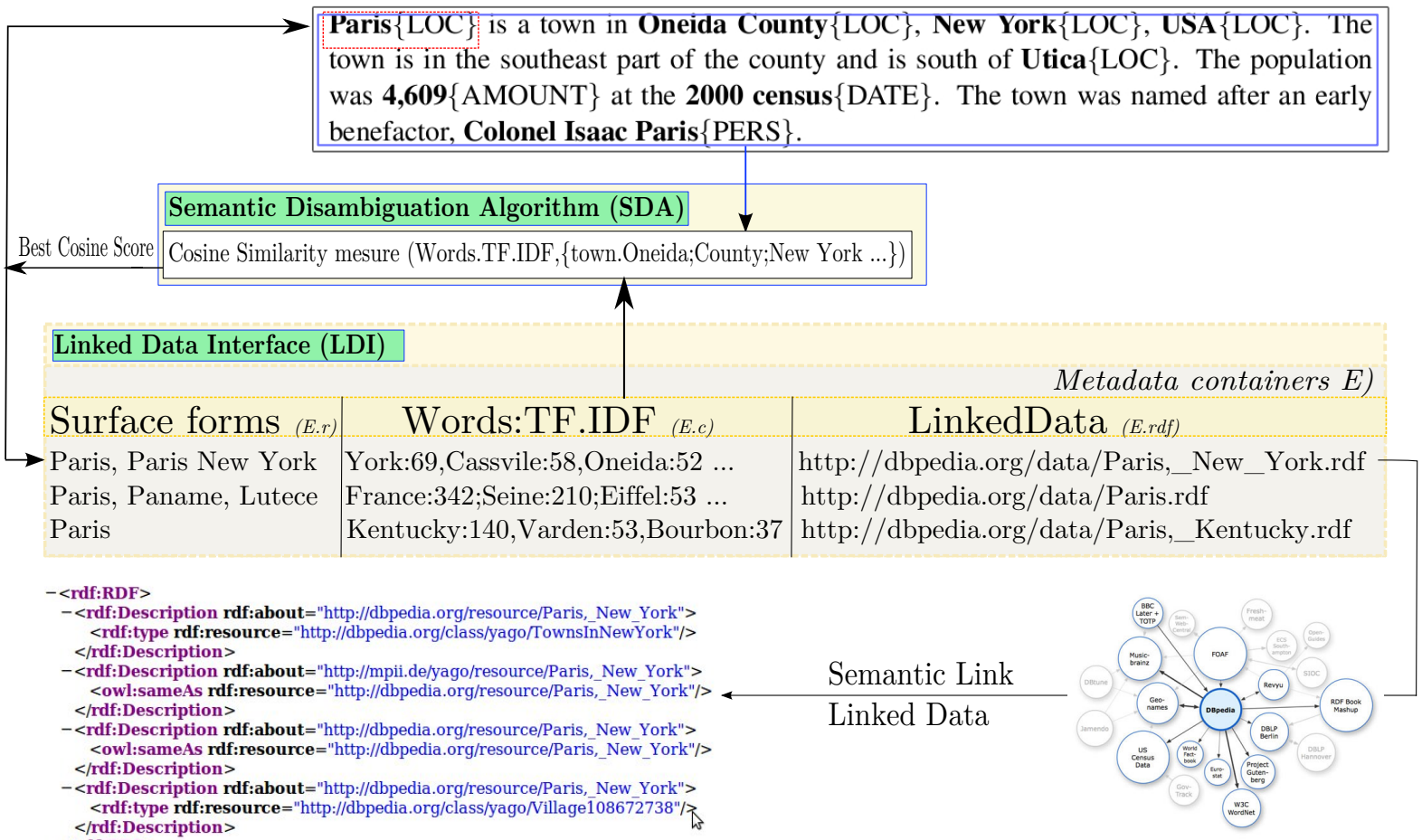
Pipeline du système

- L'étiqueteur d'entité nommées localise les EN dans le texte.
- La forme de surface de l'EN est utilisée pour localiser une ou plusieurs entités sémantiques de la LDI.
  - Une mesure de similarité est réalisée entre le contexte de l'EN et les sacs de mots de la LDI.
  - Si plus d'une candidate existe dans la LDI (ex *Paris (france)*, *Paris (Ontario)* ...), le meilleure score de similarité indique la meilleure ES.
  - Une valeur de seuil permet de rejeter les candidates à faible score (mauvaise identification présumée).

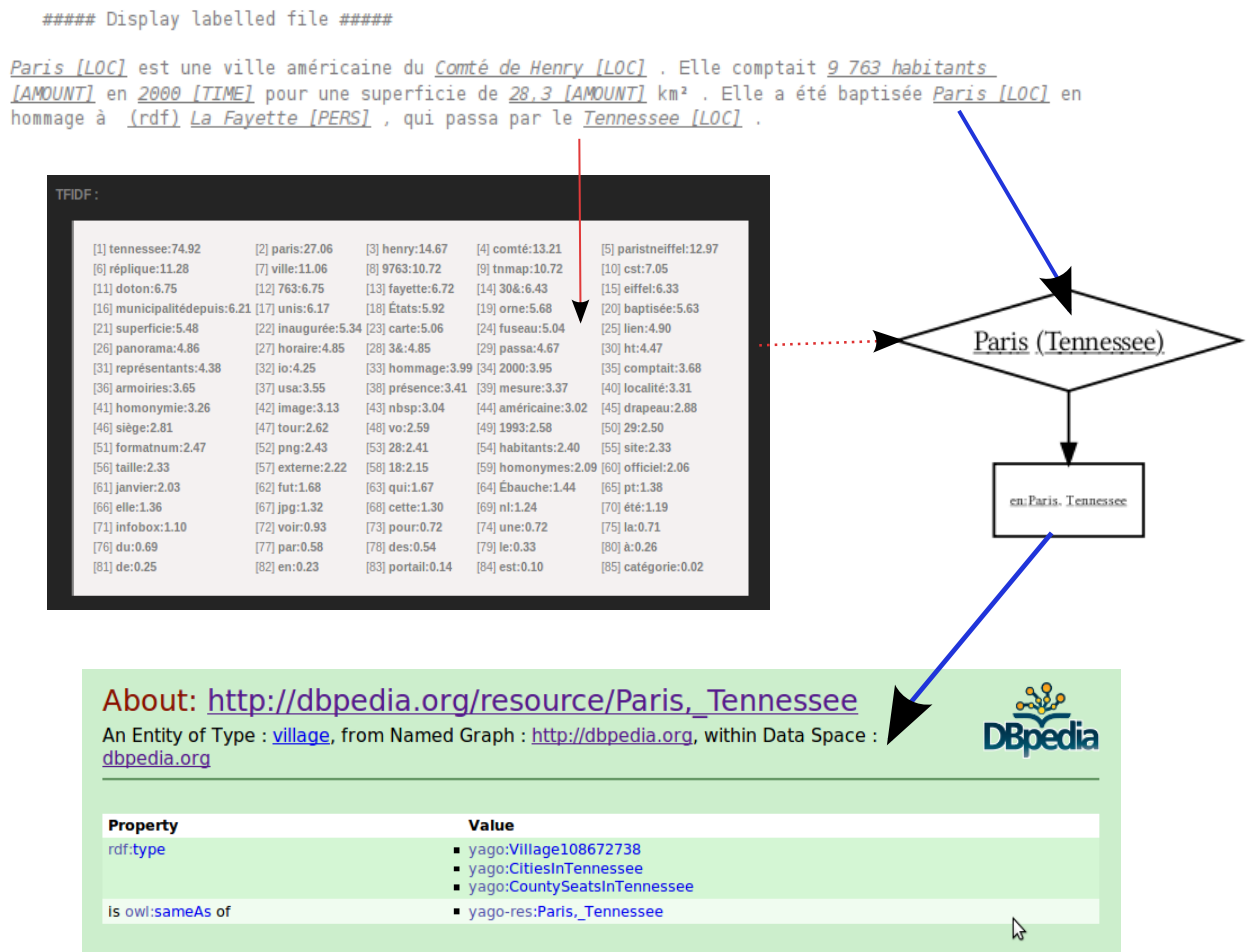
Pipeline du système

- L'étiqueteur d'entité nommées localise les EN dans le texte.
- La forme de surface de l'EN est utilisée pour localiser une ou plusieurs entités sémantiques de la LDI.
  - Une mesure de similarité est réalisée entre le contexte de l'EN et les sacs de mots de la LDI.
  - Si plus d'une candidate existe dans la LDI (ex *Paris (france)*, *Paris (Ontario)* ...), le meilleure score de similarité indique la meilleure ES.
  - Une valeur de seuil permet de rejeter les candidates à faible score (mauvaise identification présumée).
- L'instance de SE finalement retenue fournit le lien sémantique entre l'EN et son point d'entrée dans le réseau Linked Data.

Linked Data Interface (LDI)



Algorithme


Display labelled file
Paris [LOC] est une ville américaine du Comté de Henry [LOC] . Elle comptait 9 763 habitants [AMOUNT] en 2000 [TIME] pour une superficie de 28,3 [AMOUNT] km² . Elle a été baptisée Paris [LOC] en hommage à (rdf) La Fayette [PERS] , qui passa par le Tennessee [LOC] .
TFIDF :
[1] tennessee:74.92 [2] paris:27.06 [3] henry:14.67 [4] comté:13.21 [5] paristneiffel:12.97
[6] réplique:11.28 [7] ville:11.06 [8] 9763:10.72 [9] tnmap:10.72 [10] cst:7.05
[11] doton:6.75 [12] 763:6.75 [13] fayette:6.72 [14] 30&:6.43 [15] eiffel:6.33
[16] municipalitédepuis:6.21 [17] unis:6.17 [18] États:5.92 [19] orne:5.68 [20] baptisée:5.63
[21] superficie:5.48 [22] inaugurée:5.34 [23] carte:5.06 [24] fuseau:5.04 [25] lien:4.90
[26] panorama:4.86 [27] horaire:4.85 [28] 3&:4.85 [29] passa:4.67 [30] ht:4.47
[31] représentants:4.38 [32] io:4.25 [33] hommage:3.99 [34] 2000:3.95 [35] comptait:3.68
[36] armoiries:3.65 [37] usa:3.55 [38] présence:3.41 [39] mesure:3.37 [40] localité:3.31
[41] homonymie:3.26 [42] image:3.13 [43] nbsp:3.04 [44] américaine:3.02 [45] drapeau:2.88
[46] siège:2.81 [47] tour:2.62 [48] vo:2.59 [49] 1993:2.58 [50] 29:2.50
[51] formatnum:2.47 [52] png:2.43 [53] 28:2.41 [54] habitants:2.40 [55] site:2.33
[56] taille:2.33 [57] externe:2.22 [58] 18:2.15 [59] homonymes:2.09 [60] officiel:2.06
[61] janvier:2.03 [62] fut:1.68 [63] qui:1.67 [64] Ébauche:1.44 [65] pt:1.38
[66] elle:1.36 [67] jpg:1.32 [68] cette:1.30 [69] nl:1.24 [70] été:1.19
[71] infobox:1.10 [72] voir:0.93 [73] pour:0.72 [74] une:0.72 [75] la:0.71
[76] du:0.69 [77] par:0.58 [78] des:0.54 [79] le:0.33 [80] à:0.26
[81] de:0.25 [82] en:0.23 [83] portail:0.14 [84] est:0.10 [85] catégorie:0.02
Paris (Tennessee)
en:Paris, Tennessee
About: http://dbpedia.org/resource/Paris,_Tennessee
An Entity of Type : village, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org
DBpedia
Property
Value
rdf:type
yago:Village108672738
yago:CitiesInTennessee
yago:CountySeatsInTennessee
is owl:sameAs of
yago-res:Paris,_Tennessee

Expérimentations et résultats

La problématique de l'évaluation 1/3

Pourquoi évaluer?

La problématique de l'évaluation 1/3

Pourquoi évaluer?

- Besoins scientifiques (publication de résultats comparables)
- Mesure de la progression des algorithmes
- Mesure de l'évolution du système par rapport aux vocabulaires

La problématique de l'évaluation 1/3

Pourquoi évaluer?

- Besoins scientifiques (publication de résultats comparables)
- Mesure de la progression des algorithmes
- Mesure de l'évolution du système par rapport aux vocabulaires
- Anticiper l'expérience utilisateur

La problématique de l'évaluation 1/3

Pourquoi évaluer?

- Besoins scientifiques (publication de résultats comparables)
- Mesure de la progression des algorithmes
- Mesure de l'évolution du système par rapport aux vocabulaires
- Anticiper l'expérience utilisateur

Très peu de systèmes publics sont réellement évalués: combien d'erreurs ? combien d'oublis ? quels types d'erreurs ? Quel rapport au champ sémantique et lexical traité ?

La problématique de l'évaluation 2/3

La question de l'**expérience utilisateur**.

Les systèmes d'étiquetage produisent des erreurs

- Comment faire accepter ces erreurs ?
- Comment expliquer ces erreurs ?
- Comment diminuer ces erreurs ?

Solutions :

- Diminuer les erreurs en augmentant la précision.
- Faire progresser les systèmes en les entrainants et en les vérifiant régulièrement.
- Développer de nouveaux **corpus d'évaluation**.

La problématique de l'évaluation 3/3

Il n'existe pas de méthode dédiée à la mesure de la qualité d'un étiquetage compatible avec les standards du Web Sémantique.
Méthodes proches connues:

- Évaluation d'un étiquetage par classe (entités nommées, morphosyntaxe ou dépendances): précision rappel et F-mesure...
- Évaluation de chaînes de co-références (même identité d'un ensemble de mentions dans un texte): muc, blanc, etc..
- Évaluation des relations logiques (analyse sémantique, DRT): Labeled Macro F1 (CoNLL 2008).

Propositions

Comment déterminer qu'un objet textuel est correctement relié à sa ou ses représentations exactes sur le web sémantique ?

Diviser la mesure

1. Évaluer les performances de l'étiqueteur d'entités nommés: précision, rappel, F-Mesure
2. Mesurer la capacité de l'étiqueteur sémantique à établir des liens avec le réseau LinkedData
3. Évaluer un système d'étiquetage complet

Avec quel corpus de référence ?

Propositions

Cas particulier de l'étiquetage sémantique.

Ambiguïté

1. Il n'existe pas forcément une seule réponse.
2. Il peut exister une méta-réponse qui n'est pas une erreur.

1. **Meryl Linch Real Estate** est il valablement étiqueté par **Merrill Lynch** ?
2. **Manhattan** est elle une meilleure dénomination que **New-York City** pour NY downtown ?
3. **Executive director**, est il à la fois un **Senior management**, **Executive**, **Executive Officer** ?

Propositions

Toutes ces descriptions existent dans le Web Semantique par exemple via DBPedia. Mais certaines entités n'ont aucune existence sur le web sémantique !

Solution au cas particulier de l'étiquetage sémantique.

- Autoriser le système à donner plusieurs réponses correctes.
- Prévoir que le système ne donne aucune réponse.

Exemple:
Trader → `http://www.dbpedia.org/page/Trader` **OU**
`http://www.dbpedia.org/page/Stock_trader`

Corpus de test

Adaptation de corpus de référence à la tâche sémantique.

Word	POS	NE	Semantic Link
il	PRO:PER	UNK	
est	VER:pres	UNK	
20	NUM	TIME	
heures	NOM	TIME	
a	PRP	UNK	
Johannesburg	NAM	LOC.ADMI	`http://dbpedia.org/data/Johannesburg.rdf`

Table: Exemple d'annotation avec le corpus ESTER 2 NE.

Word	POS	NE	Semantic Link
Laura	NNP	PERS.HUM	*NORDF*
Colby	NNP	PERS.HUM	
in	IN	UNK	
Milan	NNP	LOC.ADMI	`http://dbpedia.org/data/Milan.rdf`

Table: Exemple d'annotation avec le sous corpus de test WSJ de CoNLL 2008.

Scores EEN

```
Wsj
-------------------------------------
amount:1465:GOOD affect:1363:total affect:1373   precision:0.99 rappel:0.93 fscore:0.96
fonc:1392:GOOD affect:972:total affect:975   precision:0.99 rappel:0.69 fscore:0.82
loc:761:GOOD affect:738:total affect:744   precision:0.99 rappel:0.96 fscore:0.98
org:1606:GOOD affect:1579:total affect:1615   precision:0.97 rappel:0.98 fscore:0.98
pers:615:GOOD affect:600:total affect:618   precision:0.97 rappel:0.97 fscore:0.97
prod:294:GOOD affect:191:total affect:212   precision:0.90 rappel:0.64 fscore:0.75
time:1018:GOOD affect:932:total affect:940   precision:0.99 rappel:0.91 fscore:0.95

precision:0.98 rappel:0.89 fscore:0.93

Ester 2
-------------------------------------
amount:237:GOOD affect:173:total affect:188   precision:0.92 rappel:0.72 fscore:0.81
fonc:425:GOOD affect:225:total affect:243   precision:0.92 rappel:0.52 fscore:0.67
loc:1309:GOOD affect:1145:total affect:1417   precision:0.80 rappel:0.87 fscore:0.84
org:1341:GOOD affect:790:total affect:842   precision:0.93 rappel:0.58 fscore:0.72
pers:1138:GOOD affect:1045:total affect:1099   precision:0.95 rappel:0.91 fscore:0.93
prod:59:GOOD affect:15:total affect:26   precision:0.57 rappel:0.254 fscore:0.35
time:1046:GOOD affect:680:total affect:702   precision:0.96 rappel:0.65 fscore:0.77

precision:0.90 rappel:0.73 fscore:0.80
```

Scores liens sémantiques

```
-------------- Sur le premier rang -----------
By entity  :    Ok:1255:        Bad:177 (1432)
By instance:    Ok:3738:        bad:715
Précision=83.94

-------------- Avec 3 Nbest -----------------
By entity  :    Ok:1347         Bad:53
By instance:    4221:ok         Bad:232
Précision=94.79
-------------------------------------------

Details by semantic category:
-- Good semtag --
FONC:999 /ORG:1248 /PROD:233 /LOC:639 /CONCEPT:75 /PERS:544 /
Total:3738

-- By Category: Bad semtag --
FONC:176 /ORG:308 /UNK:3 /PROD:11 /LOC:118 /PERS:62 /CONCEPT:37
Total:715
```

Conclusions et perspectives

Conclusions.

- Nous avons présenté un système capable d'ajouter un lien sémantique à des entités contenues dans un texte libre.
- Les ressources sémantiques sont standards, de type URI, compatibles avec le web sémantique via le réseau *Linked Data*.
- Nous avons introduit le principe de la *Linked Data Interface*.

Nos évaluations montrent que notre système parvient à établir un lien fiable dans plus de 86% des cas

Travaux futurs.

Renforcer la détection de l'identité précise d'un mot ou d'une séquence de mot en utilisant des méthodes complémentaires.

Détection de chaînes de Co-Références

- *Poly-co, an unsupervised co-reference detection system* (INLG, GREC 2010)
- *Poly-co, a Perceptron approach for co-reference detection* (CoNLL Shared Task 2011)

Introduire les informations sur la relation logique des éléments textuels.

Analyse syntaxique et sémantique

- Étiquetage des dépendances
- Découvertes des relations sémantiques

Merci.

Et venez l'essayer sur www.wikimeta.org !