Annotation automatique de texte avec le web sémantique et le réseau Linked Data.

Eric Charton, Michel Gagnon, Benoit Ozell

ÉCOLE POLYTECHNIQUE MONTRÉAL

Plan

Le Web sémantique et le réseau Linked Data

Le Web Sémantique

Le Web sémantique désigne un ensemble de technologies visant à rendre le contenu des ressources du World Wide Web accessible et utilisable par les programmes et agents logiciels, grâce à un système de métadonnées formelles, utilisant notamment la famille de langages développés par le W3C.

Le Web Sémantique

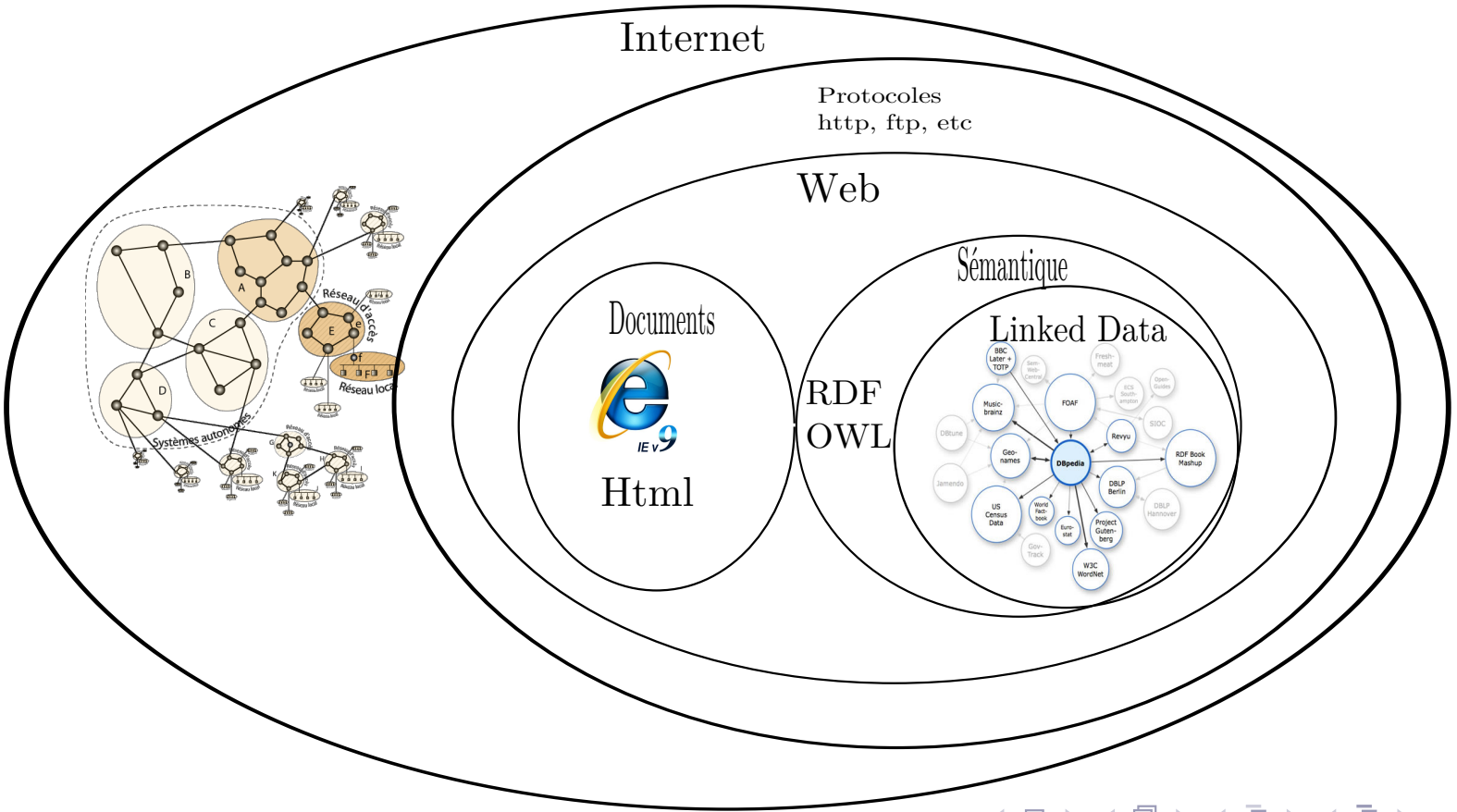
Du Web des documents au Web Sémantique

- Le Web sémantique est entièrement fondé sur le Web.
- Le Web sémantique utilise les protocoles du Web (http,URI, XML).
- Le Web sémantique introduit ses propres standards (RDF, OWL, SPARQL).

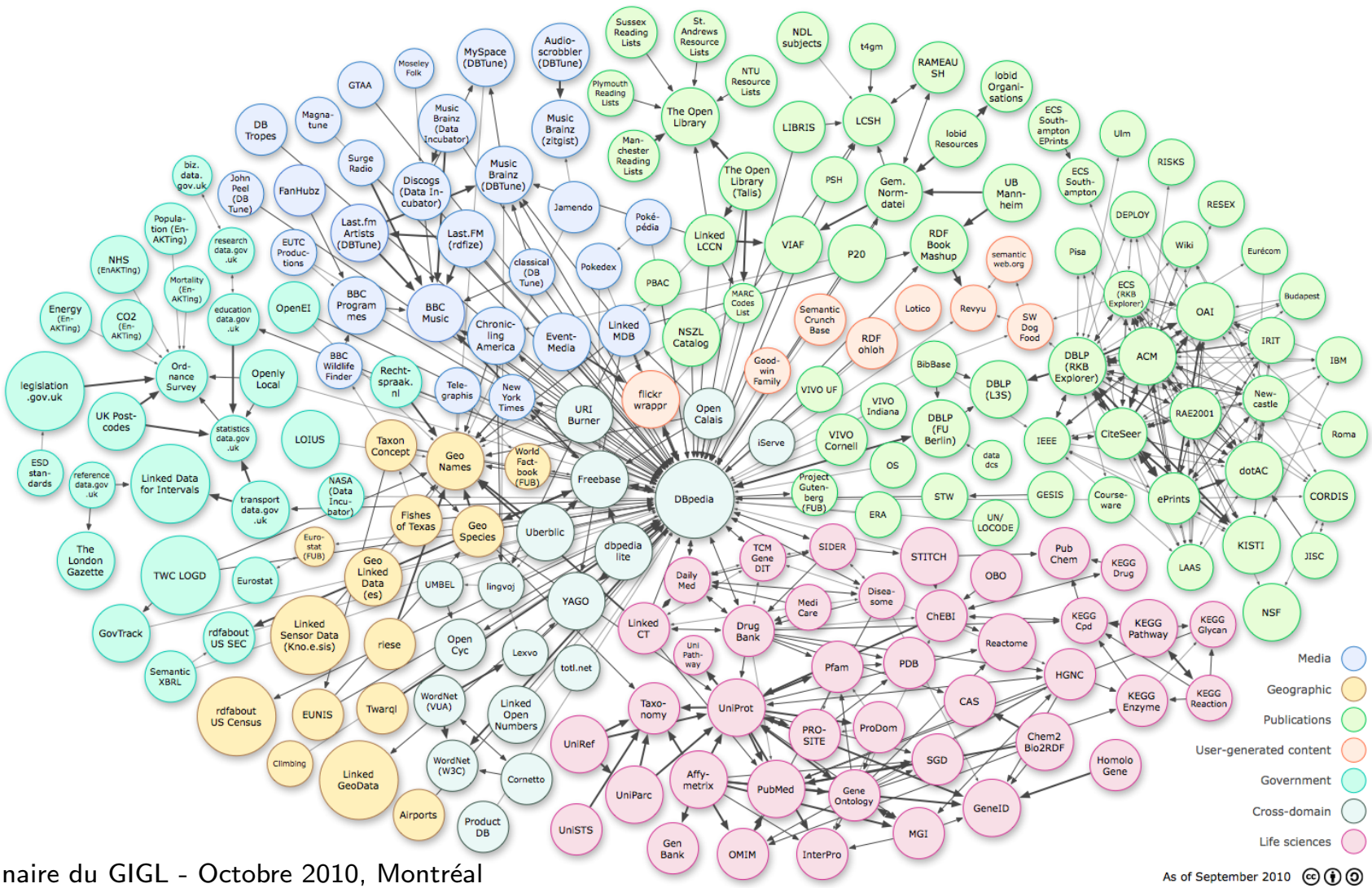
L'avancement du Web sémantique dans le monde est suivi par le W3C dans le cadre d'un projet Semantic Web Advanced Deployment (SWAD).

Le réseau Linked Data

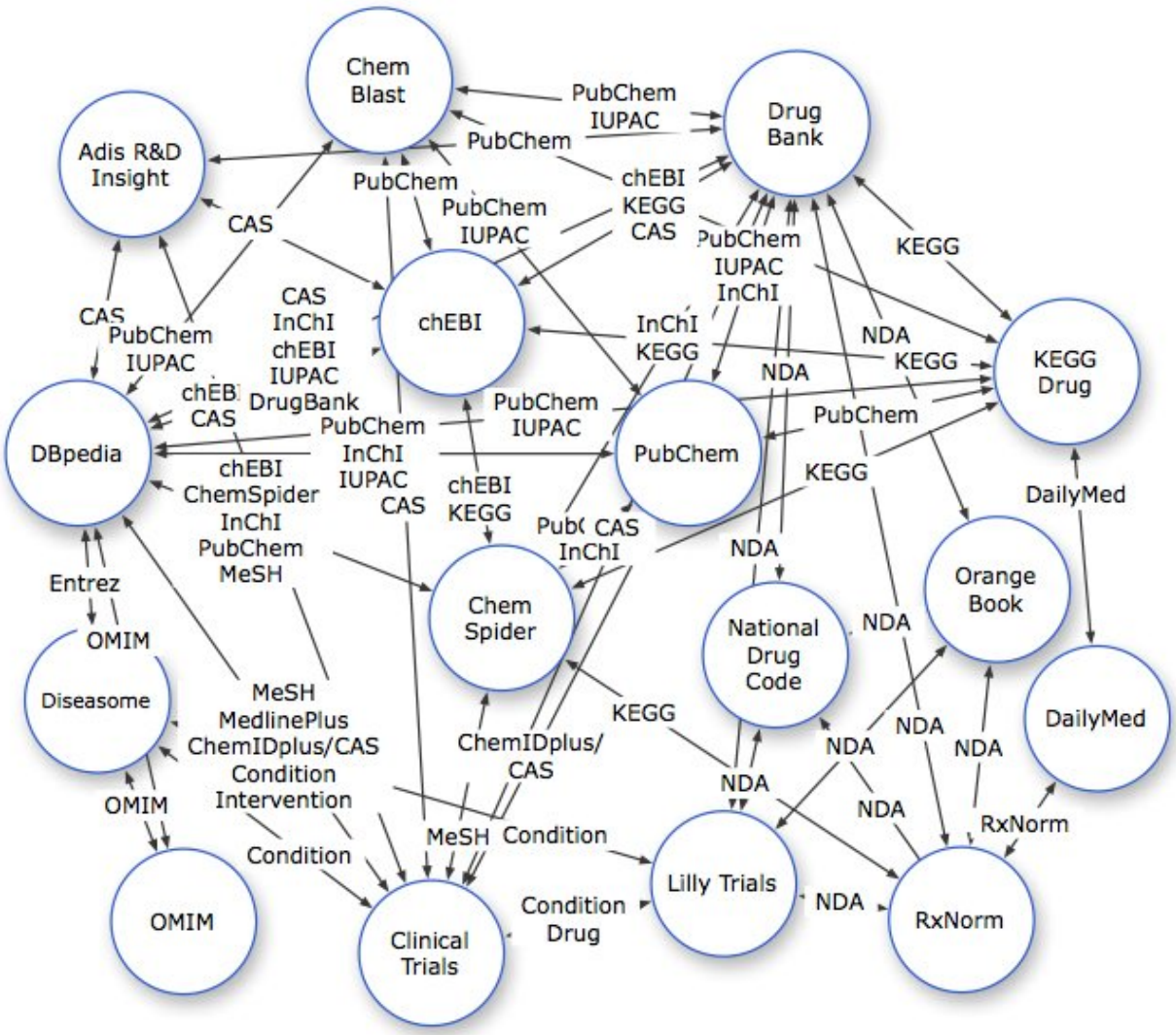
Linked Data est un sous ensemble du Web sémantique. Il décrit un ensemble de règles et de méthodes pour publier, partager et connecter des données sémantiques sur le Web.



Le réseau Linked Data (en septembre 2010)



Le réseau Linked Data : interconnexion



Le réseau Linked Data : une croissance hors norme

Des milliards de données inter-connectées. Une croissance extrêmement rapide.

Des données facile à produire..

- Transformation de bases de données (Libris, US Census).
- Extraction de connaissance structurées (DBpédia, Yago).
- Publications de données gouvernementales (data.Gov.Uk).

Le réseau Linked Data: pour quoi faire ?

Des données difficiles à utiliser.

- Le Web des documents est exploré par des méthodes de recherche fonctionnelles (algorithmes de Recherche d'Information - RI).
- Mais comment explorer un Web des données et pour quoi faire ?
  - Comment analyser les données (raisonnement)
  - Comment relier les données (étiquetage).
  - Comment obtenir une information (SQR).

Le réseau Linked Data: pour quoi faire ?

Des données difficiles à utiliser.

- Le Web des documents est exploré par des méthodes de recherche fonctionnelles (algorithmes de Recherche d'Information - RI).
- Mais comment explorer un Web des données et pour quoi faire ?
  - Comment analyser les données (raisonnement)
  - Comment relier les données (étiquetage).
  - Comment obtenir une information (SQR).

Proposition d'application

Utiliser le réseau Linked Data dans une tâche d'étiquetage de texte.

Problématique de l'étiquetage sémantique

Qu'est ce qu'une entité dans un texte ? Quel est son rôle ?

Texte sans entité nommée

- Une personne ivre a percuté en voiture celle d'une autre personne.

Qu'est ce qu'une entité dans un texte ? Quel est son rôle ?

Texte sans entité nommée

- Une personne ivre a percuté en voiture celle d'une autre personne.

Texte avec entité nommée

- George W Bush ivre a percuté en Humer la Lincoln de Barack Obama.

Qu'est ce qu'une entité dans un texte ? Quel est son rôle ?

Texte sans entité nommée

- Une personne ivre a percuté en voiture celle d'une autre personne.

Texte avec entité nommée

- George W Bush ivre a percuté en Humer la Lincoln de Barack Obama.

L'entité nommée (EN) introduit un degré supplémentaire d'information et donne une identité aux protagonistes.

Nature particulière des Entités Nommées et des Entités Sémantiques

La mise en relation de connaissances avec des mots d'un texte peut être vue comme le prolongement de la tâche d'étiquetage d'EN.
Il existe pourtant une différence essentielle entre l' EN dans son texte et sa représentation ontologique:

Limitation sémantique des Entités Nommées

- L'étiquetage d'EN revient à affecter un taxon (pers.hum, loc.fac, org.com) à un mot ou un groupe de mot.
- Ce taxon (ou label de classe) ne permet pas de refléter toutes les particularités sémantiques de l'entité.

Exemple: (1) localiser des entités

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

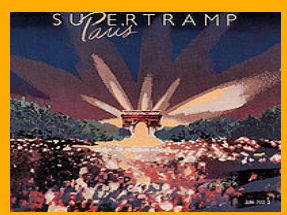
Exemple: (2) désambiguïser la classe

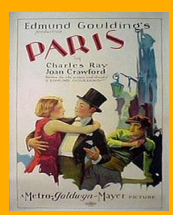
La position de Paris à un carrefour entre les itinéraires commerciaux terrestres et fluviaux au cœur d'une riche région agricole en a fait une des principales villes de France au cours du Xe siècle, avec des palais royaux, de riches abbayes et une cathédrale

Détection des entités nommées
Étiquetage

?





Un astéroïde (3317 Paris), une ville, un paquebot, un disque, un film?

Exemple: (3) affecter une 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 des entités nommées
Étiquetage : une ville LOC.ADMI

Exemple: (4) il subsiste une ambiguïté ...

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
Étiquetage : une ville LOC.ADMI

Quelle Ville ?

Une classe d'EN ne permet pas de gérer cette ambiguïté

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
Étiquetage : une ville LOC.ADMI


Paris (Ontario)


Paris


Paris (Tenessee)

Paris (Idaho)


Paris (Maine)


Paris (Kentucky)

L'étiquetage sémantique

Relier une Entité Nommée à une description ontologique

L'étiquetage sémantique

Relier une Entité Nommée à une description ontologique

- Étiqueter une Entité Nommée selon un processus classique

L'étiquetage sémantique

Relier une Entité Nommée à une description ontologique

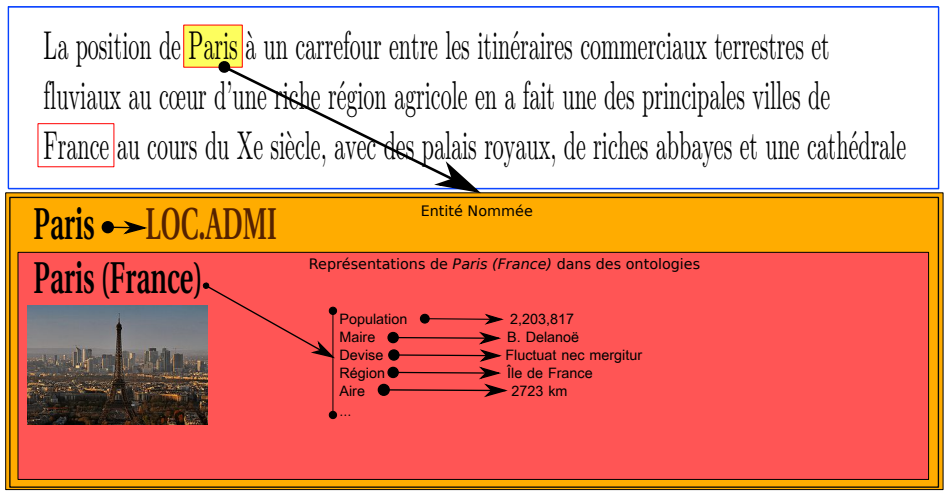
- Étiqueter une Entité Nommée selon un processus classique
- Associer à l'Entité détectée un lien vers un graphe qui décrit son identité Sémantique

L'étiquetage sémantique

Relier une Entité Nommée à une description ontologique

- Étiqueter une Entité Nommée selon un processus classique
- Associer à l'Entité détectée un lien vers un graphe qui décrit son identité Sémantique

Revient à localiser une EN puis à la relier à sa représentation ontologique exacte, par exemple issue du réseau Linked Data.



Principes applicatifs d'un étiqueteur sémantique

Difficulté: deux méthodes applicatives opposées

Étiquetage d'une Entité Nommée

Étiquetage sémantique

Difficulté: deux méthodes applicatives opposées

Étiquetage d'une Entité Nommée

- Fait appel à un classifieur numérique (CRF, SVM) qui infère d'après le contexte, en fonction d'un apprentissage préalable.

Étiquetage sémantique

Difficulté: deux méthodes applicatives opposées

Étiquetage d'une Entité Nommée

- Fait appel à un classifieur numérique (CRF, SVM) qui infère d'après le contexte, en fonction d'un apprentissage préalable.
- Prévu pour généraliser la reconnaissance d'un nombre de classe limité (4/250 selon les normes taxonomiques) d'après un contexte textuel "générique".

Étiquetage sémantique

Difficulté: deux méthodes applicatives opposées

Étiquetage d'une Entité Nommée

- Fait appel à un classifieur numérique (CRF, SVM) qui infère d'après le contexte, en fonction d'un apprentissage préalable.
- Prévu pour généraliser la reconnaissance d'un nombre de classe limité (4/250 selon les normes taxonomiques) d'après un contexte textuel "générique".

(ex une société ORG.COM) → mots du contexte → *CAC40*, *Bilan*, *Effectifs*, *chiffre d'affaire* etc)..

Étiquetage sémantique

Difficulté: deux méthodes applicatives opposées

Étiquetage d'une Entité Nommée

- Fait appel à un classifieur numérique (CRF, SVM) qui infère d'après le contexte, en fonction d'un apprentissage préalable.
- Prévu pour généraliser la reconnaissance d'un nombre de classe limité (4/250 selon les normes taxonomiques) d'après un contexte textuel "générique".

(ex une société ORG.COM) → mots du contexte → *CAC40*, *Bilan*, *Effectifs*, *chiffre d'affaire* etc)..

Étiquetage sémantique

- Mise en relation d'une Entité Nommée avec un graphe

Difficulté: deux méthodes applicatives opposées

Étiquetage d'une Entité Nommée

- Fait appel à un classifieur numérique (CRF, SVM) qui infère d'après le contexte, en fonction d'un apprentissage préalable.
- Prévu pour généraliser la reconnaissance d'un nombre de classe limité (4/250 selon les normes taxonomiques) d'après un contexte textuel "générique".

(ex une société ORG.COM) → mots du contexte → *CAC40*, *Bilan*, *Effectifs*, *chiffre d'affaire* etc)..

Étiquetage sémantique

- Mise en relation d'une Entité Nommée avec un graphe
- Nombre de graphe illimité: égal au nombre d'entités uniques à identifier. La généralisation et l'approche numérique classique deviennent inutilisables.

Difficulté: deux méthodes applicatives opposées

Étiquetage d'une Entité Nommée

- Fait appel à un classifieur numérique (CRF, SVM) qui infère d'après le contexte, en fonction d'un apprentissage préalable.
- Prévu pour généraliser la reconnaissance d'un nombre de classe limité (4/250 selon les normes taxonomiques) d'après un contexte textuel "générique".

(ex une société ORG.COM) → mots du contexte → *CAC40*, *Bilan*, *Effectifs*, *chiffre d'affaire* etc)..

Étiquetage sémantique

- Mise en relation d'une Entité Nommée avec un graphe
- Nombre de graphe illimité: égal au nombre d'entités uniques à identifier. La généralisation et l'approche numérique classique deviennent inutilisables.
- Fait appel à la nature sémantique identifiée d'après le contexte. A chaque Entité Nommée unique correspond un contexte non généralisable

Difficulté: deux méthodes applicatives opposées

Étiquetage d'une Entité Nommée

- Fait appel à un classifieur numérique (CRF, SVM) qui infère d'après le contexte, en fonction d'un apprentissage préalable.
- Prévu pour généraliser la reconnaissance d'un nombre de classe limité (4/250 selon les normes taxonomiques) d'après un contexte textuel "générique".

(ex une société ORG.COM) → mots du contexte → *CAC40*, *Bilan*, *Effectifs*, *chiffre d'affaire* etc)..

Étiquetage sémantique

- Mise en relation d'une Entité Nommée avec un graphe
- Nombre de graphe illimité: égal au nombre d'entités uniques à identifier. La généralisation et l'approche numérique classique deviennent inutilisables.
- Fait appel à la nature sémantique identifiée d'après le contexte. A chaque Entité Nommée unique correspond un contexte non généralisable

(ex Paris (France) → mots du contexte → *Seine*, *Tour Eiffel* etc).

Proposition

Séparer les deux tâches d'EEN et d'ES en introduisant une représentation intermédiaire qui assure le lien entre une Entité Nommée dans son texte et sa représentation ontologique.

Proposition

Séparer les deux tâches d'EEN et d'ES en introduisant une représentation intermédiaire qui assure le lien entre une Entité Nommée dans son texte et sa représentation ontologique.

Linked Data Interface (LDI)

Proposition

Séparer les deux tâches d'EEN et d'ES en introduisant une représentation intermédiaire qui assure le lien entre une Entité Nommée dans son texte et sa représentation ontologique.

Linked Data Interface (LDI)

- La **Linked Data Interface (LDI)** est à la fois statistique et conceptuelle.

Proposition

Séparer les deux tâches d'EEN et d'ES en introduisant une représentation intermédiaire qui assure le lien entre une Entité Nommée dans son texte et sa représentation ontologique.

Linked Data Interface (LDI)

- La **Linked Data Interface (LDI)** est à la fois statistique et conceptuelle.
- Pour chaque ES elle modélise un contexte possible composé d'un sac de mots avec ses poids TF.IDF.

Proposition

Séparer les deux tâches d'EEN et d'ES en introduisant une représentation intermédiaire qui assure le lien entre une Entité Nommée dans son texte et sa représentation ontologique.

Linked Data Interface (LDI)

- La **Linked Data Interface (LDI)** est à la fois statistique et conceptuelle.
- Pour chaque ES elle modélise un contexte possible composé d'un sac de mots avec ses poids TF.IDF.
- Pour chaque ES elle inclut un ou plusieurs liens vers une instance ontologique.

Ressources de construction de la **Linked Data Interface (LDI)**

Construite avec les ressources du Web

- 900000 Entités Sémantiques en Français, 3 M en Anglais (personnes, lieux, produits, organisations) issues de Wikipédia.
- Chaque fiche encyclopédique fournit les mots contextuels et poids TF.IDF correspondant à des Entités Sémantiques.
- Chaque Entité Sémantique issue de Wikipédia peut être reliée à une ou plus représentations ontologique (DBPédia, CIA World Factbook ...) par le réseau Linked Data grâce aux tables de correspondances.

Architecture du système

Étapes de détection et de mise en relation

- Un étiqueteur d'Entité Nommées détecte les EN dans le texte.

Architecture du système

Étapes de détection et de mise en relation

- Un étiqueteur d'Entité Nommées détecte les EN dans le texte.
- L'instance correspondant à l'Entité Nommée détectée est recherchée dans la Linked Data Interface

Architecture du système

Étapes de détection et de mise en relation

- Un étiqueteur d'Entité Nommées détecte les EN dans le texte.
- L'instance correspondant à l'Entité Nommée détectée est recherchée dans la Linked Data Interface
  - Une mesure de **similarité cosinus** est réalisé entre un vecteur représentatif des poids des mots du contexte de l'entité et ceux des candidats des instances de l'ontologie de liaison.

Architecture du système

Étapes de détection et de mise en relation

- Un étiqueteur d'Entité Nommées détecte les EN dans le texte.
- L'instance correspondant à l'Entité Nommée détectée est recherchée dans la Linked Data Interface
  - Une mesure de **similarité cosinus** est réalisé entre un vecteur représentatif des poids des mots du contexte de l'entité et ceux des candidats des instances de l'ontologie de liaison.
  - Si plusieurs candidates (ex *Paris (france)*, *Paris (Ontario)* ...), le meilleurs score de similarité détermine la bonne instance.

Architecture du système

Étapes de détection et de mise en relation

- Un étiqueteur d'Entité Nommées détecte les EN dans le texte.
- L'instance correspondant à l'Entité Nommée détectée est recherchée dans la Linked Data Interface
  - Une mesure de **similarité cosinus** est réalisé entre un vecteur représentatif des poids des mots du contexte de l'entité et ceux des candidats des instances de l'ontologie de liaison.
  - Si plusieurs candidates (ex *Paris (france)*, *Paris (Ontario)* ...), le meilleurs score de similarité détermine la bonne instance.
  - Une constante de seuil permet de rejeter les entités candidates peu fiables.

Architecture du système

Étapes de détection et de mise en relation

- Un étiqueteur d'Entité Nommées détecte les EN dans le texte.
- L'instance correspondant à l'Entité Nommée détectée est recherchée dans la Linked Data Interface
  - Une mesure de **similarité cosinus** est réalisé entre un vecteur représentatif des poids des mots du contexte de l'entité et ceux des candidats des instances de l'ontologie de liaison.
  - Si plusieurs candidates (ex *Paris (france)*, *Paris (Ontario)* ...), le meilleurs score de similarité détermine la bonne instance.
  - Une constante de seuil permet de rejeter les entités candidates peu fiables.
- L'instance de l'ontologie de liaison fournit le lien vers la représentation sémantique de l'entité dans l'ontologie.

Linked Data Interface (LDI)

Paris{LOC} is a town in **Oneida County**{LOC}, **New York**{LOC}, **USA**{LOC}. The town is in the southeast part of the county and is south of **Utica**{LOC}. The population was **4,609**{AMOUNT} at the **2000 census**{DATE}. The town was named after an early benefactor, **Colonel Isaac Paris**{PERS}.

Semantic Disambiguation Algorithm (SDA)

Best Cosine Score — Cosine Similarity mesure (Words.TF.IDF,{town.Oneida;County;New York ...})

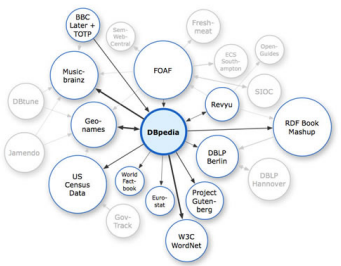
Linked Data Interface (LDI)

Metadata containers E)

Surface forms *(E.r)*	Words:TF.IDF *(E.c)*	LinkedData *(E.rdf)*
Paris, Paris New York	York:69,Cassvile:58,Oneida:52 ...	http://dbpedia.org/data/Paris,_New_York.rdf
Paris, Paname, Lutece	France:342;Seine:210;Eiffel:53 ...	http://dbpedia.org/data/Paris.rdf
Paris	Kentucky:140,Varden:53,Bourbon:37	http://dbpedia.org/data/Paris,_Kentucky.rdf

```
-<rdf:RDF>
  -<rdf:Description rdf:about="http://dbpedia.org/resource/Paris,_New_York">
     <rdf:type rdf:resource="http://dbpedia.org/class/yago/TownsInNewYork"/>
   </rdf:Description>
  -<rdf:Description rdf:about="http://mpii.de/yago/resource/Paris,_New_York">
     <owl:sameAs rdf:resource="http://dbpedia.org/resource/Paris,_New_York"/>
   </rdf:Description>
  -<rdf:Description rdf:about="http://dbpedia.org/resource/Paris,_New_York">
     <owl:sameAs rdf:resource="http://dbpedia.org/resource/Paris,_New_York"/>
   </rdf:Description>
  -<rdf:Description rdf:about="http://dbpedia.org/resource/Paris,_New_York">
     <rdf:type rdf:resource="http://dbpedia.org/class/yago/Village108672738"/>
   </rdf:Description>
</rdf:RDF>
```

Semantic Link
Linked Data



Expériences et résultats

Méthode d'évaluation

Associer à des EN détectées dans un fichier de référence (Gold Standard) un lien vers leurs *instances* contenues dans une *ontologie descriptive*. C'est l'établissement de ce lien qui caractérise le principe de l'*étiquetage sémantique*.
On utilise un corpus de référence français (ESTER 2) et anglais (CoNLL 2008).

Exemple de Corpus d'évaluation annoté: ESTER 2

Mot	POS	NE	Lien sémantique
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 du corpus français de test de la campagne ESTER 2. Un lien vers DBpedia correspondant à une EN est ajouté dans une nouvelle colonne.

Exemple de Corpus d'évaluation annoté: CoNLL

Mot	POS	NE	Lien sémantique
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 du corpus de test anglais CoNLL 2008. L'annotation sémantique *NORDF* est utilisée quand aucun lien RDF n'est disponible.

Caractéristiques de couverture des Corpus de test

La couverture des *métadonnées* du LDI n'est pas complète. Une partie seulement des EN du corpus WSJ CoNLL 2008 et de ESTER 2 bénéficie d'une entrée dans le réseau Linked Data.

	ESTER 2 2009 (Français)			WSJ CoNLL 2008 (Anglais)		
Labels	Entités du corpus de test	Entités équivalentes dans LDI	Couverture (%)	Entités du corpus de test	Entités équivalentes dans LDI	Couverture (%)
PERS	1096	483	44%	612	380	62%
ORG	1204	764	63%	1698	1129	66%
LOC	1218	1017	83%	739	709	96 %
PROD	59	23	39%	61	60	98 %
GPE						
Total	3577	2287	64%	3110	2278	73%

Capacité du système à associer correctement une identité sémantique à une Entité Nommée déjà détectée par un système d'EEN.

	Français	ESTER	2		Anglais	CoNLL		
NE	[no α]	Recall	[α]	Recall	[no α]	Recall	[α]	Recall
PERS	483	0.96	1096	0.91	380	0.93	612	0.94
ORG	764	0.91	1204	0.90	1129	0.85	1608	0.86
LOC	1017	0.94	1218	0.92	709	0.84	739	0.82
PROD	23	0.60	59	0.50	60	0.85	61	0.85
Total	2287	0.93	3577	0.9	2278	0.86	3020	0.86

Table: Résultats de l'étiqueteur sémantique appliqué aux corpus de test ESTER 2 et WSJ CoNLL 2008.

Conclusions

Conclusions.

- Nous avons présenté un système d'étiquetage sémantique (SE) utilisable pour compléter un système d'étiquetage par Entités Nommées (EEN).
- Une mesure de similarité cosinus a été mise en œuvre pour établir un lien entre une Entité Nommée (EN) étiquetée dans sa phrase et sa représentation ontologique standardisée dans la ressource DBpedia.
- Cette méthode nous a permis d'associer à 94% des Entités Nommées (EN) du corpus d'évaluation de la campagne ESTER 2, connues dans Wikipédia, un lien vers leur description ontologique dans DBpedia.
- L'interface LDI est compatible avec le Web sémantique et exhaustive.

Perspectives

Insertion dans le projet Gitan.

- Utiliser l'étiquetage pour déterminer l'identité des concepts contenus dans un texte.

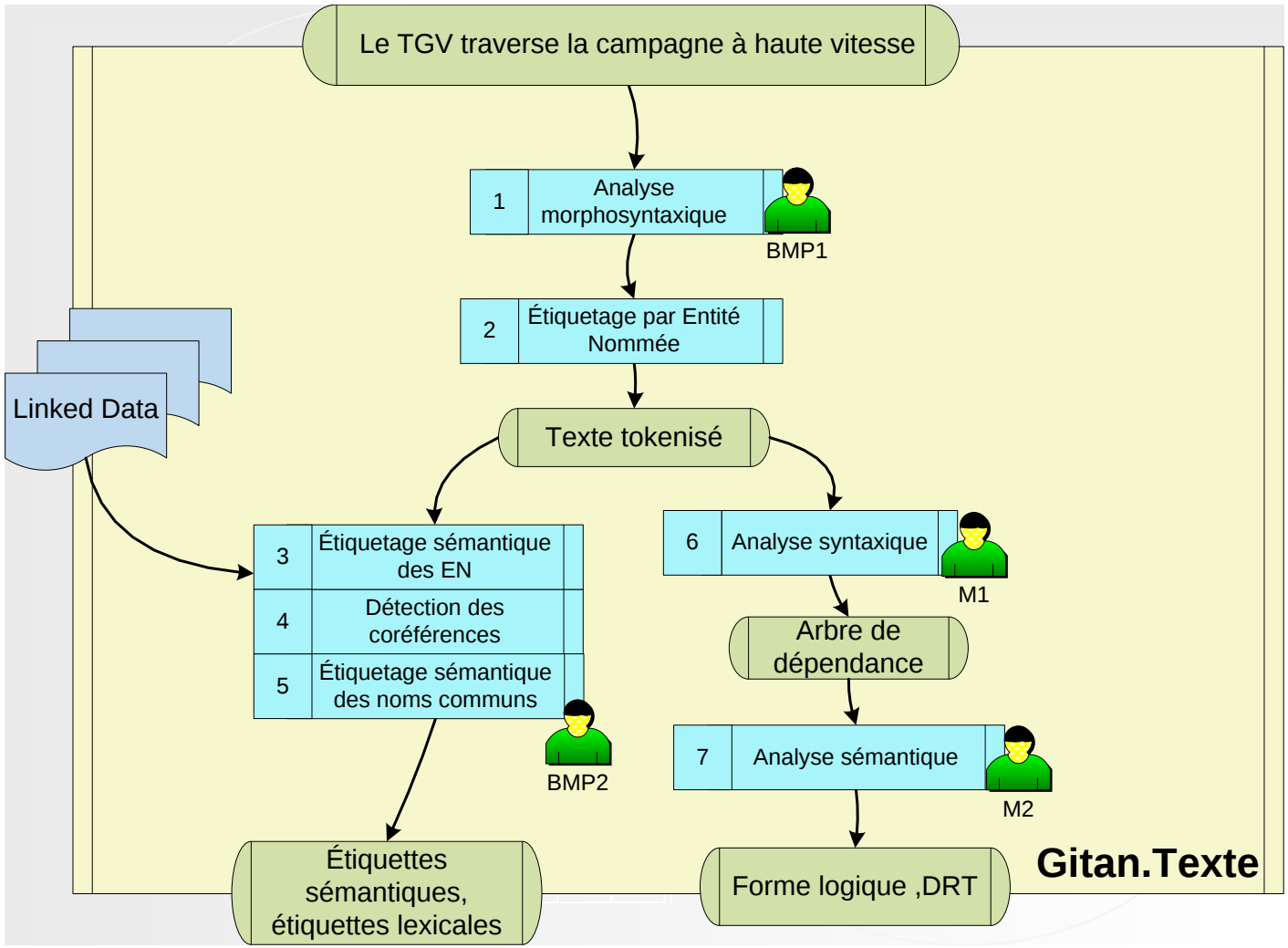
Exemple

Le TGV traverse la campagne à haute vitesse.

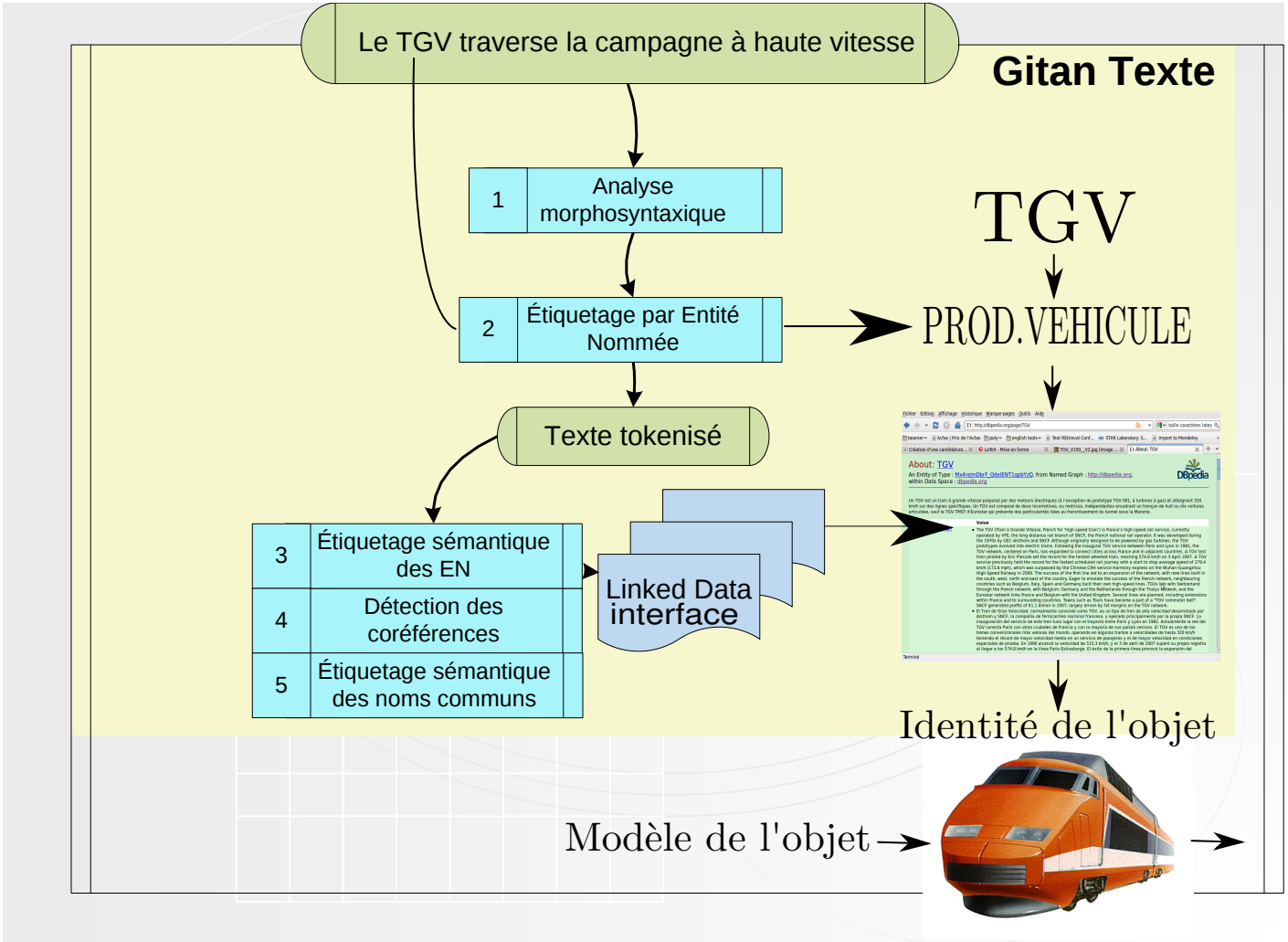
Application

Extraction du concept TGV pour le représenter.

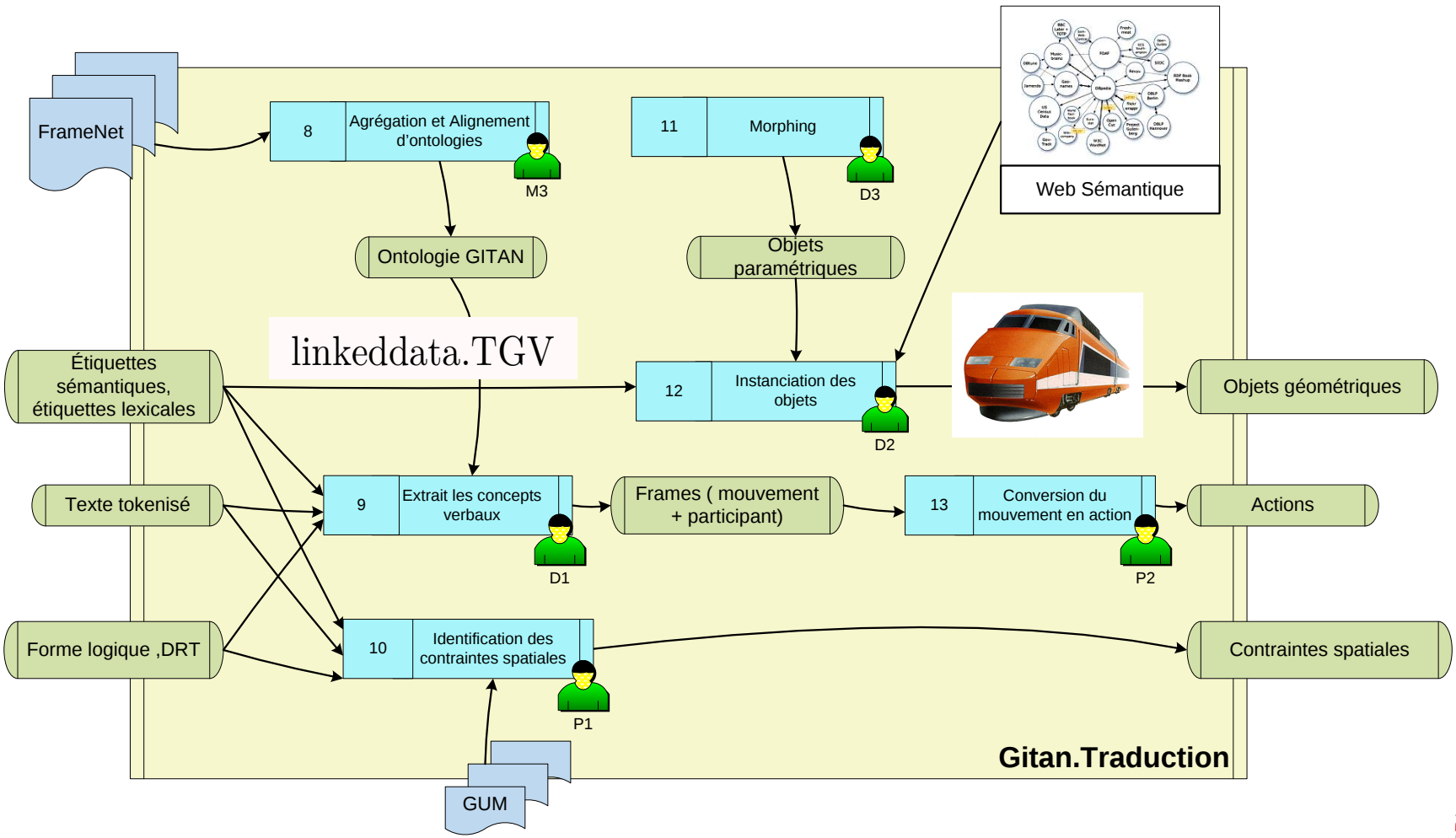
Insertion dans le projet Gitan: chaîne de traitement TAL



Insertion dans le projet Gitan: identité de l'objet



Insertion dans le projet Gitan: choix d'un modèle.



Et bientôt ...

Quelques références

- Photo du “Linking Open Data cloud diagram“ par Richard Cyganiak et Anja Jentzsch. http://lod-cloud.net/
- ”Extension d’un système d’étiquetage d’entités nommées en étiqueteur sémantique“, TALN2010, Charton E., Gagnon M., Ozell B. (première publication sur les principes de base)

Merci.

Ambiguïté insoluble !

Paris [loc.admi] est une ville américaine du comté de Henry [loc.admi] (Tennessee [loc.admi]) . Elle comptait 9763 habitants [AMOUNT] en 2000 [TIME] pour une superficie de 28,3 km [AMOUNT] . Elle a été baptisée Paris [loc.admi] en hommage à La Fayette [pers.hum] , qui passa par le Tennessee [loc.admi] La réplique de la Tour Eiffel [loc.fac] qui orne la ville fut inaugurée le 29 janvier 1993 [TIME] , en présence de représentants de la ville de Paris [loc.admi] Cette réplique mesure 18,30 m [AMOUNT] .

Réseau de formes de surface