SAIMSI
Suivi Adaptatif Interlingue et Multisource des Informations

Christian Fluhr
GEOLSemantics
christian.fluhr@geolsemantics.com

34 mois
263 p/mois
3,2 M€

Comité de pilotage

labellisé par le pôle

Le consortium

Les partenaires et leur rôle:

GEOLSemantics est le leader du projet , développe le text mining interlingue, des bases de données textuelles et du traitement de la parole

Cassidian est l'architecte du système basé sur la plate-forme **Weblab**. Il est aussi l'intégrateur des différents modules et en fournit certains

Mondeca est en charge de la construction de la base de connaissance, des raisonnements automatiques, de la gestion et de l'utilisation de la base de connaissance

Le LIP6 est en charge des technologies de reconnaissance de l'auteur ainsi que des méthodes d'évaluation

L'IREENAT est en charge des aspects juridiques et déontologiques du projet SAIMSI

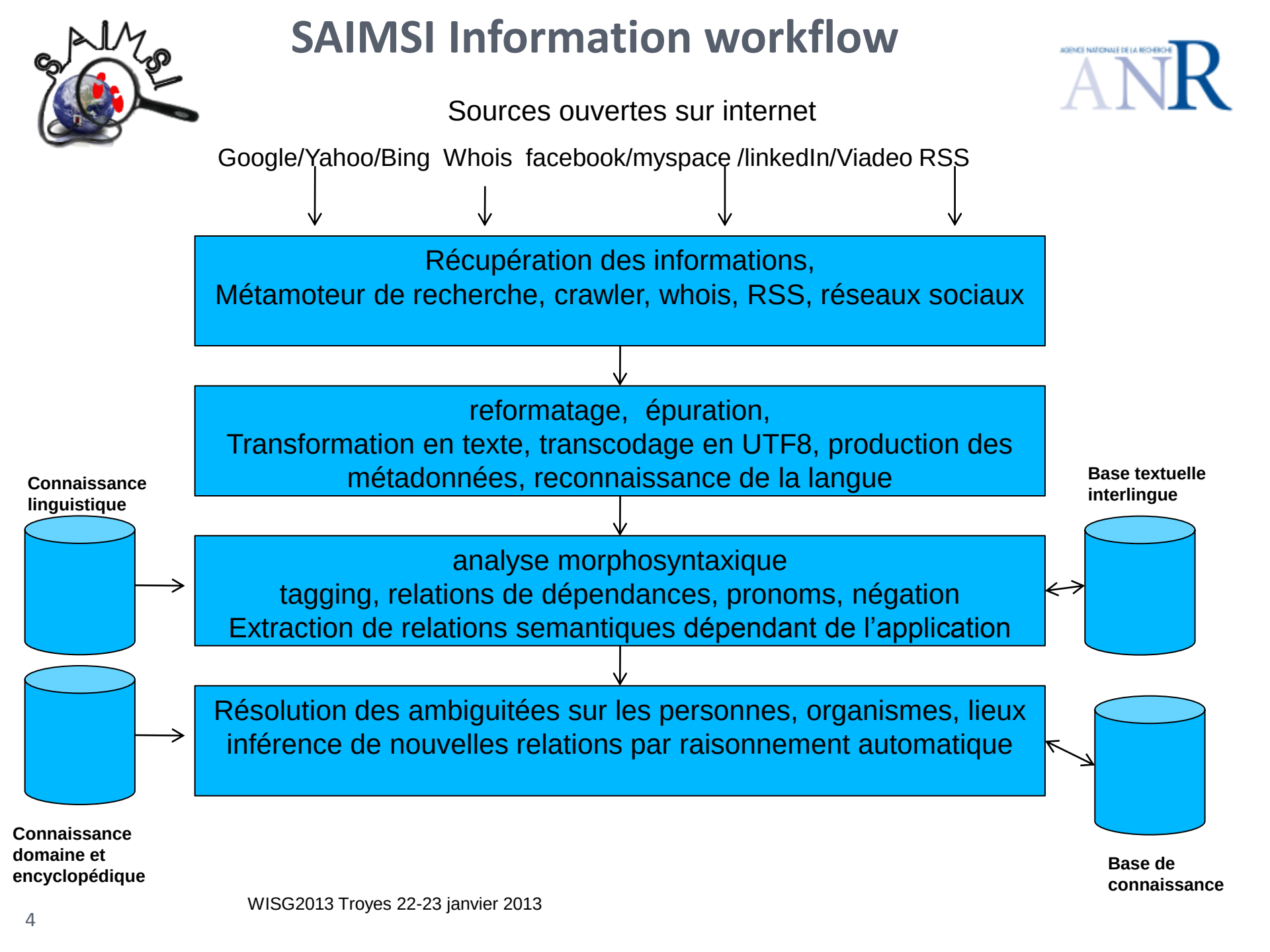
Partenariat : biométrie vocale d'Agnitio, transcription de parole Vocapia research

LES OBJECTIFS DU PROJET

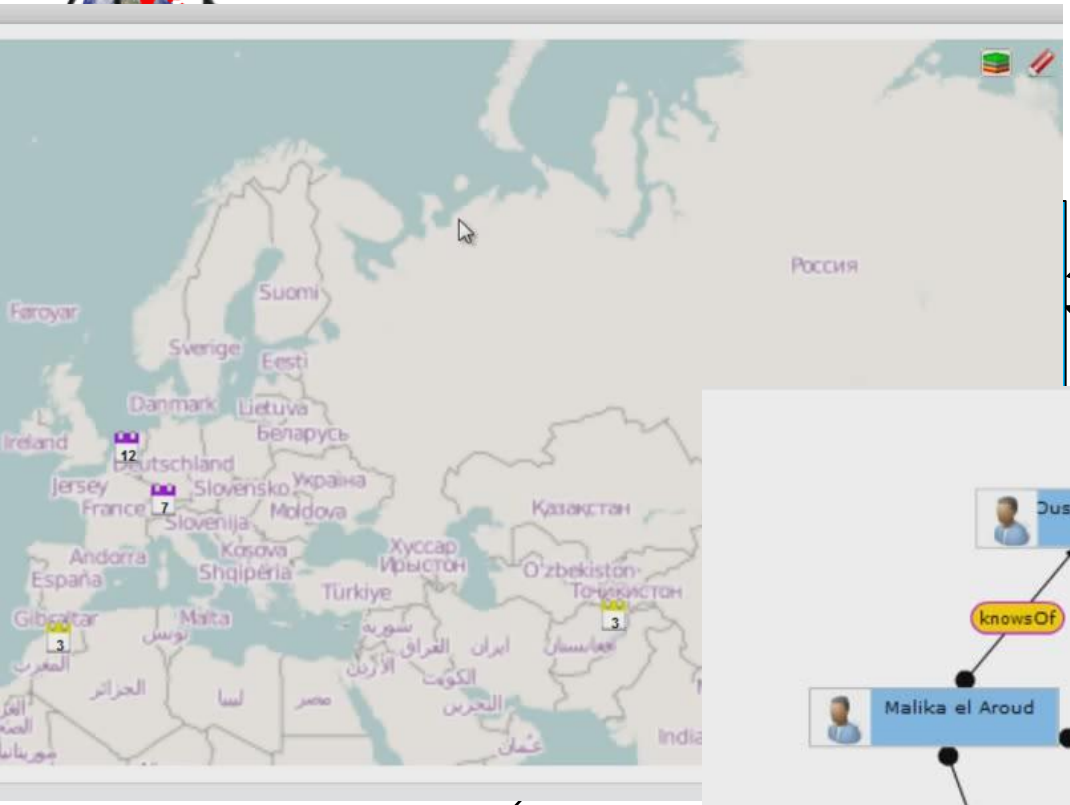
Suivi des activités de personnes soupçonnées d'activités illicites (terrorisme, drogue, blanchiment d'argent, …),

- à partir de sources ouvertes sur internet : sites webs, presse, réseaux sociaux
- en plusieurs langues (français, anglais, arabe et chinois)
- en deux média texte et parole
- traitement des données en flux

SAIMSI Information workflow



Fonctions de SAIMSI

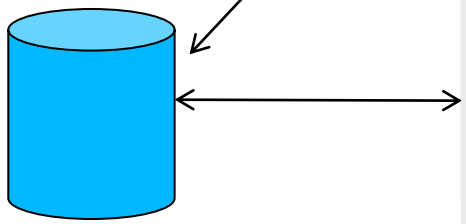


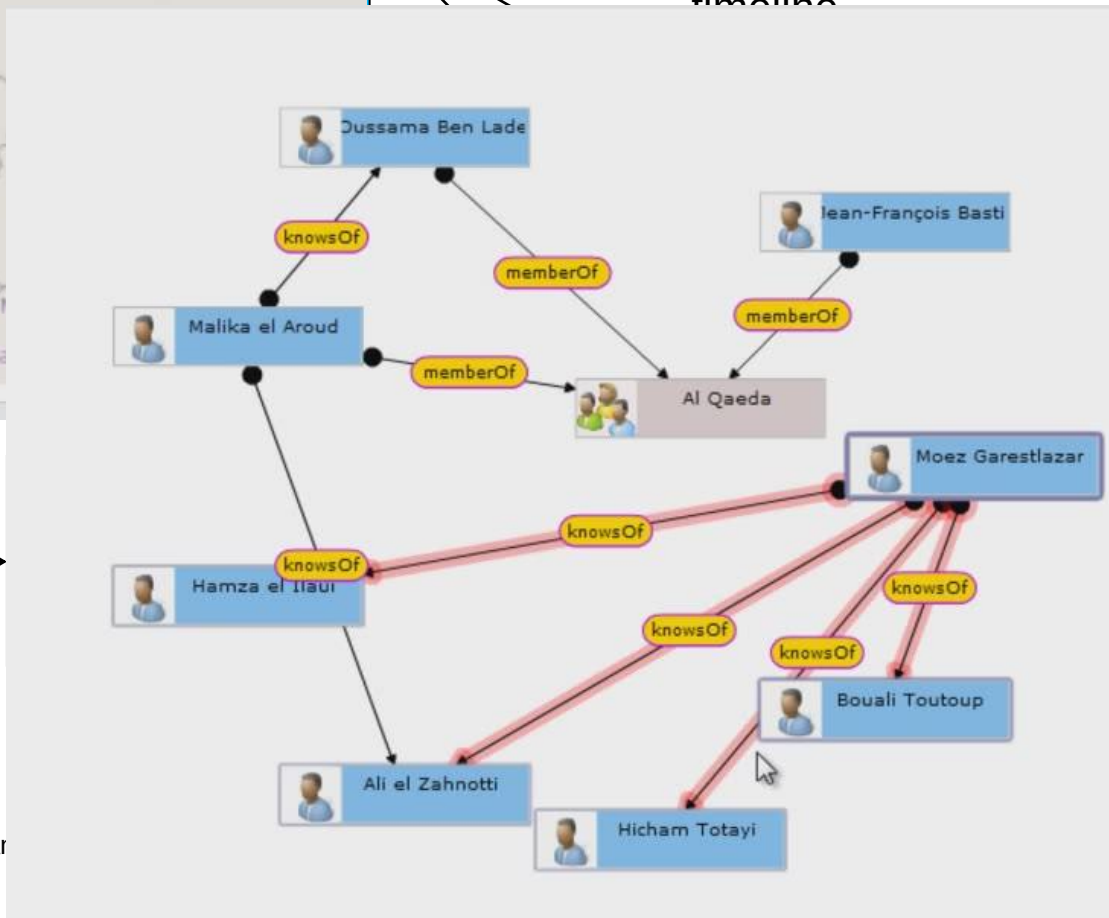
requête

Carte géographiques

timeline

interlingue





Les challenges (verrous)

- Réalisation d'une ontologie large satisfaisant à la fois les besoins des utilisateurs et les contraintes de l'extraction automatique à partir des textes

- Extraction de connaissances basée sur une analyse morphosyntaxique multilingue généraliste profonde et une extraction sémantique liée à l'ontologie de l'application

- Représentation des connaissances indépendantes de la langue source des documents

- Identification de l'auteur d'un texte

- Traitement des textes au fur et à mesure

Les deux cas d'utilisation retenus

Premier cas d'utilisation:
Prescripteur : DCRI
Sujet : les activités jihadistes
Les langues principales concernées: français, anglais, chinois et arabe

La CNIL a autorisé l'utilisation de recherches sur seulement Malika el Aroud et Oussama Ben Laden

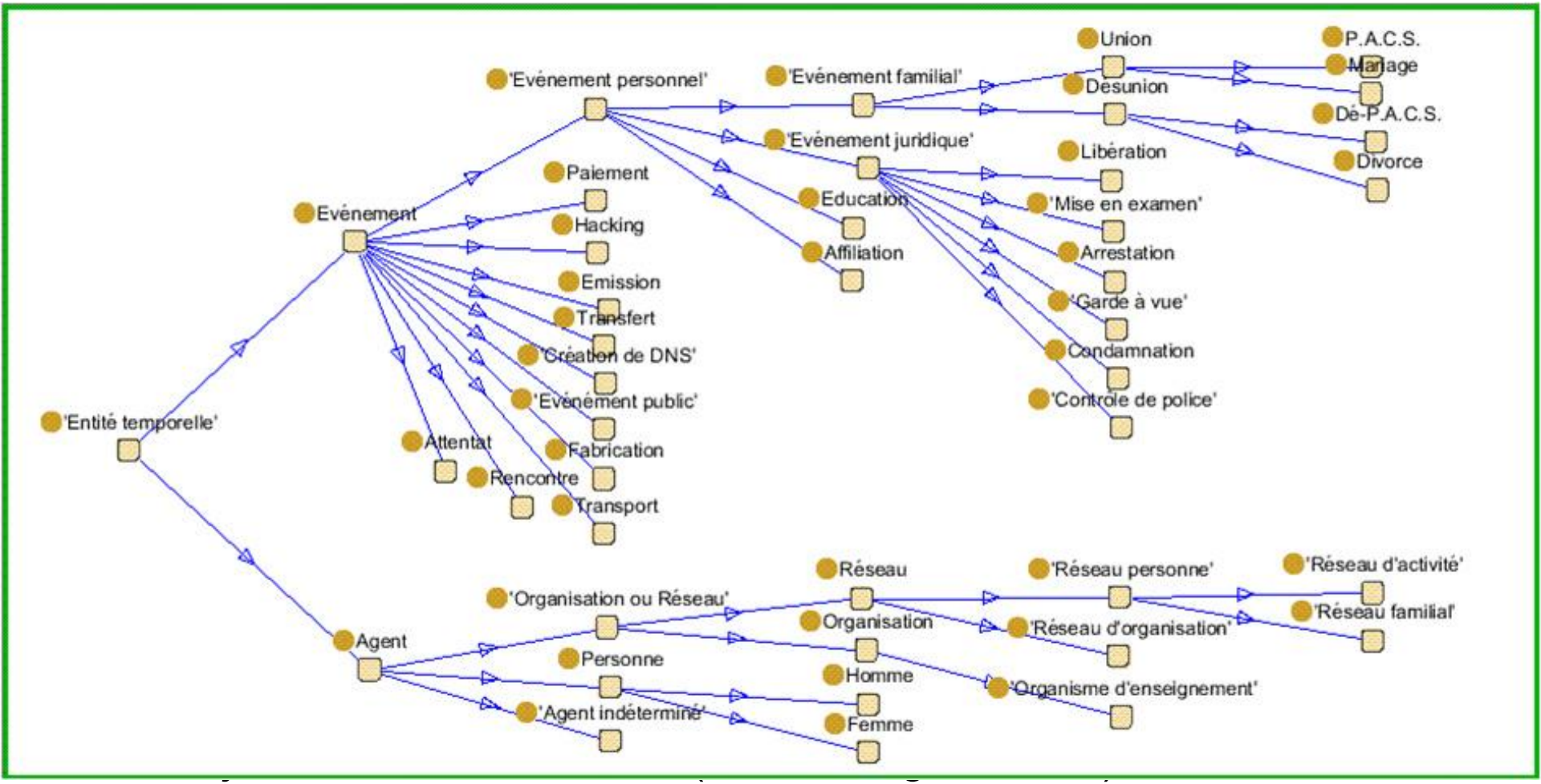
DRM intéressée par les fiches biographiques

Deuxième cas d'utilisation:
Prescripteur: OCLCTIC
Sujet: le phishing et les malwares
Langues: français, anglais, russe et chinois

Arrêt des développements à cause de la limitation à deux noms par la CNIL.

Une ontologie de la sécurité

Décrit ce qui peut être cherché à propos de personnes:



rôles dans un organisme, relation comme étudiant dans un organisme d'enseignement.

Les étapes de la constitution de la base de connaissance

Choix d'une stratégie basée sur une analyse morphosyntaxique profonde généraliste

avantages:

- minimiser les résultats faux

- trouver des informations rares qui ne peuvent être rattrapées par la redondance (signaux faibles)

- permet plus facilement de passer d'une application à l'autre

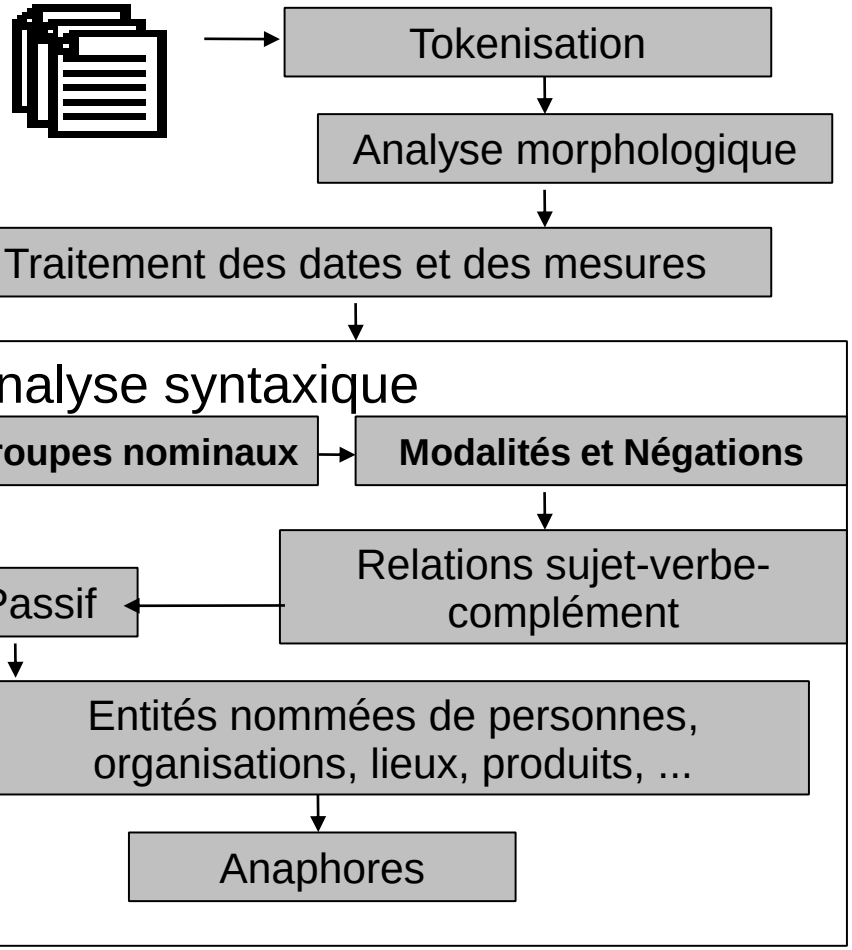
Inconvénients:

- demande plus d'effort de développement pour minimiser les informations non reconnues

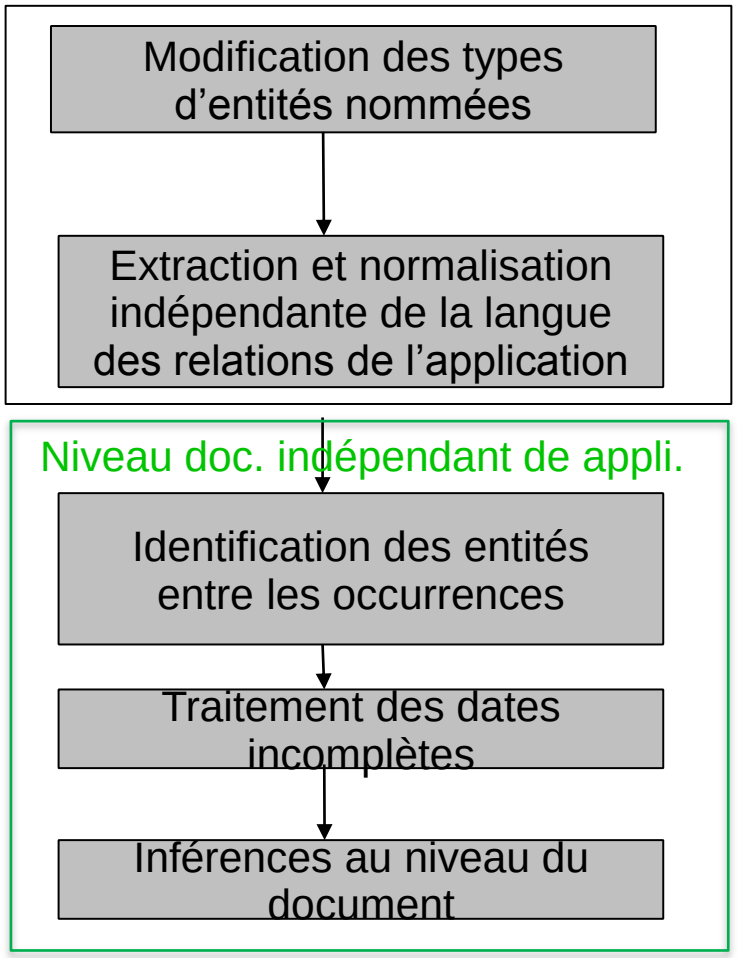
- traitement plus lourds

L'extraction de connaissance

Analyse morpho-syntaxique profonde multilingue indépendante de l'application



Extraction d'information dépendant de l'application



Extraction au niveau de la phrase

Exemple: dépêche du 13 mai 2012

« Hollande se rendra mardi à Berlin. Il rencontrera la chancelière Angela Merkel. »

L'interprétation de la première phrase est simple à partir du résultat de l'analyse morphosyntaxique
« Hollande » est l'agent de l'action « se rendre ». « mardi » est la date et« Berlin » est le lieu.

Les règles d'extraction doivent tester deux hypothèses pour l'action se rendre (aller ou se constituer prisonnier). Du fait de la présence d'une destination « Berlin » c'est la première hypothèse qui est retenue.

"Hollande" est considéré par l'analyse morphosyntaxique comme un lieu mais la règle de reconnaissance d'un déplacement ne peut avoir comme agent qu'une personne ou une organisation. En l'absence de certitude on peut le qualifier d'agent. En fait, on introduit « Hollande » aussi comme personne pour éviter l'ambiguïté.

Extraction au niveau de la phrase

Exemple: dépêche du 13 mai 2012

« Hollande se rendra mardi à Berlin. Il rencontrera la chancelière Angela Merkel. »

La deuxième phrase pose plus de problèmes.

On perd beaucoup d'informations dans un texte si on ne traite pas les pronoms. Il est nécessaire ici d'identifier il à Hollande. C'est ce que fait l'analyse au niveau du paragraphe et même au niveau du document par exemple dans les biographies.

Extraction au niveau du document

Exemple: dépêche du 13 mai 2012

« Hollande se rendra mardi à Berlin. Il rencontrera la chancelière Angela Merkel. »

Dans la première phrase il y a une date incomplète.
Les dates incomplètes sont traitées en fonction de la date de publication ou d'une date citée avant dans le document. Les dates incomplètes sont traduites en format ISO.
Mardi= 20120515

Toutes les dates sont données comme un intervalle d'incertitude. Par exemple *2011* sera traduit par Datebeg=20110101 Dateend=20111231,
La semaine prochaine si la date de publication est le 13 mai 2012:
Datebeg=20120514 Dateend=20120520

La deuxième phrase ne précise pas où et quand la rencontre va avoir lieu. En l'absence d'une indication contraire de l'auteur, un déplacement suivi d'une rencontre laisse penser que la date et le lieu sont communs entre les deux phrases.

Extraction au niveau du document

La même personne peut être citée à différentes occurrences dans le texte en utilisant une étiquette différente.
Exemple: Nicolas Sarkozy ….. Le président Sarkozy …..Sarkozy

Les différentes actions ou attributs doivent être liés à une URI unique même si la personne est nommée de manières différentes.

En cas de types de différentes spécificités, le plus spécifique est attribué. Si Hollande dans un cas est agent et dans un autre est personne on le qualifiera de personne.

On peut bien sûr avoir dans le même texte plusieurs personnes différentes de même nom de famille. Dans ce cas l'auteur précise le prénom ou un titre discriminant.

Ex: Abdul Aziz Bastin est le fils de Jean-François Bastin.

Ces considérations s'appliquent aussi aux lieux, organisations et produits.

Normalisation des informations provenant de plusieurs langues

La langue de représentation des connaissances est l'anglais.

Les valeurs des attributs doivent être transformées en anglais.

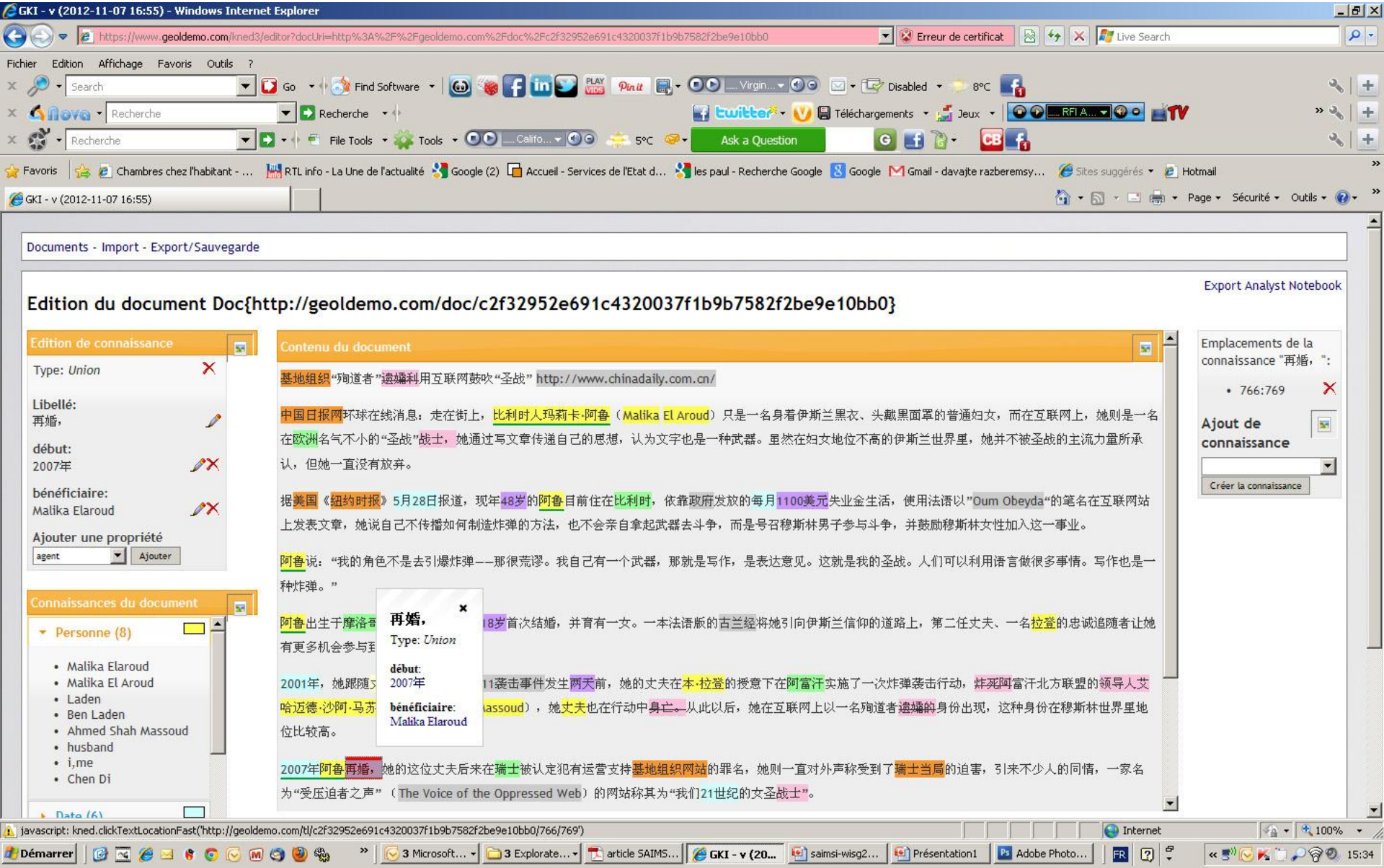
Cela pose des problèmes de listes d'autorité multilingues qui peuvent être très longues comme par exemple les noms de métiers.

Certaines valeurs ont une traduction ambiguë en anglais. Par exemple belle-fille qui peut être « daughter-in-law » ou « stepdaughter ».

Les noms propres de personnes ou de lieux exprimés dans des langues avec un autre jeu de caractères comme l'arabe et le chinois doivent être exprimés en écriture latine pour être comparables à leur forme dans des textes européens.

On utilise les dictionnaires pour les noms connus et le translittérateur pour les autres noms. Exemple 胡锦涛 → Hú Jǐntāo → Hu Jintao

EXEMPLES DE CONNAISSANCE EXTRAITE D'UN DOCUMENT



AJOUT DES INFORMATIONS DANS LA BASE DE CONNAISSANCE

La principale difficulté est de déterminer si une personne identifiée dans le nouveau document est ou non une des personnes déjà répertoriée dans la base de connaissance

En cas de même nom,

certains critères permettent de dire que ce sont deux personnes différentes: date de naissance, lieu de naissance incompatibles, différence de patronymes ou de suffixe (Jr Sr)

certains critères permettent de dire qu'il s'agit de la même personne: téléphone, mail, adresse

AJOUT DES INFORMATIONS DANS LA BASE DE CONNAISSANCE

Certaines personnes de noms différents sont la même personne.

C'est en particulier le cas de variantes de latinisation de noms arabes, russes ou chinois

Pour cela on produit des hypothèses des orthographications possibles à l'aide d'un translittérateur

Exemple: pour Oussama ben laden
osama bin ledan, osama bin Laden, ozama ban ladin, osama ben ladane, osama ben ladin, asama binladdan, osama binleden, osama bin lden, osama ben ledan, …

SATURATION DES CONNAISSANCES DE LA BASE

Les connaissances issues des textes ne donnent qu'une vision partielle de la réalité telle que peut l'appréhender un être humain

Il est possible **d'inférer de nouvelles relations ou attributs** par un raisonnement automatique

Par exemple si « Dominique est la sœur de Jean » On va pouvoir déduire de la relation de fratrie fournie par le traitement linguistique que Dominique est de sexe féminin.

D'autre part cette relation doit être complétée par Jean est le frère de Dominique si on veut connaître toutes les relations familiales de Jean

IDENTIFICATION DE L'AUTEUR D'UN TEXTE

L'identification de l'auteur d'un texte est un problème difficile.

Les méthodes classiques s'appuient sur des traits assez simples de surface des textes: trigrammes de lettres, mots, succession de catégories grammaticales et sur des classifieurs simples.

La méthode évaluée par le LIP6 (feature bagging) consiste à faire un apprentissage de classifieurs sur un ensemble aléatoire d'ensemble de traits et ensuite de faire un vote

Évaluation PAN 2012

Méthode	Accuracy
SVM sur 3000 traits	71,6
Bagging 600 classifieurs avec chacun 100 traits	79,4

LES DÉBOUCHÉS DE SAIMSI

Aide à la productivité dans l'alimentation manuelle de bases de données structurées comme le système I2 d'IBM par exemple en traitant les procès verbaux d'enquêtes

Exemple: graphe IBM I2 Analyst Notebook des relations de Malika el Aroud

Adaptation de l'ontologie pour traiter le domaine économique: Rachats de sociétés, nominations de dirigeants, déclarations de dirigeants

Applications pour la presse qui n'a pas les mêmes contraintes juridiques pour les bases contenant desnoms de personnes

ÉVOLUTIONS DES TECHNOLOGIES

Optimisation des temps de traitement pour se rapprocher des besoins du big data (en cours).

Traitement différencié par type de documents:

biographies, CV, dépêches de presse

réseaux sociaux

Extraction de relations liées à la structure du document

Développement d'un raisonneur capable de mieux prendre en compte des incertitudes de l'extraction à partir du langage naturel

Merci

Questions?