Chantier d’usage ISTEX

Inria-Alpage

Équipe

Aazhar

Patrice Lopez
(kermitt2)

Laurent Romary

Les enjeux de la fouille de textes scientifiques

- e-science : exploitation du calcul intensif pour tirer profit de volumes massifs de données scientifiques
  - ➡ biologie, bioinformatique, génomie, physique, astronomie, sciences humaines

- Pour (Hey, 2009) : le début d'une quatrième révolution scientifique où l'usage des ordinateurs permettra la création de nouveaux concepts, idées, modèle et simulations, et une "renaissance" scientifique

- Fouille de textes scientifiques : rendre exploitable l'information contenu dans la littérature scientifique et technique
  - ➡ Amélioration des outils de recherche d'information,
  - ➡ Construction de bases de connaissances,
  - ➡ À plus long terme, production automatique d'hypothèses scientifiques (Evans, 2010)

Les blocages

- Droit d'accès à la littérature scientifique et technique
- Légalité de la fouille de texte (US/UK/JP vs FR/DE)
- Besoin de couverture, données à jour (+ 1,5M articles/an)

GROBID

- Difficulté d'exploitation du format PDF / pauvreté et incohérence des metadonnées
- "*the information that I can extract from an article, at least for me, is not quite the information I want*", Shreejoy Tripathy (neuroscientifique)

Les défis scientifiques et techniques

- Gestion de la volumétrie : million de documents, milliards d'annotations, graphe avec milliard de noeuds
- Gestion de l'incertitude et du bruit: les meilleures techniques de fouille de textes font des erreurs, beaucoup d'erreurs...
- *Machine learning* : domaines et scénarios plus complexes et ouverts que le web marchand et les "add clicks"
- Prise en compte des méthodologies, opinions, réseaux, etc. propres à chaque discipline

Tirer profit de la fouille de textes dans ISTEX

- Un très grand volume de données afin de neutraliser le bruits, les erreurs d'extractions, etc.
- Décloisonner les disciplines: croiser des champs/ disciplines
- Travail de normalisation, nettoyage des données extraites
- Capturer le langage spécialisé

Objectifs du chantier d'usage ISTEX

- Tester et passer à l'échelle nos propres annotateurs
- Combiner analyse sémantique multi-domaine et extraction spécialisée de nomenclature scientifique
- Monter en charge notre infrastructure d'analyse documentaire scientifique
- Illustrer l'intérêt d'ISTEX pour la fouille de texte

Travail du chantier d'usage ISTEX

1. **Annotations spécialisées en information scientifique et technique sur un grand échantillon -** 10 à 20% de l'ensemble du corpus ISTEX
   - (N)ERD : entités connus / catégorisation
   - BIO-GROBID/BEAST : biotech
   - CHEMICAL-GROBID : chimie
   - GROBID-quantities : mesures physiques
2. Encodage TEI stand-off des annotations
3. Démonstrateur exploitant ces annotations basé sur notre infrastructure anHALytics

Approche

- Les approches par apprentissages automatiques dominent toutes les compétitions en extraction d'information de textes scientifiques

How to improve the performance of a ML application?

- Get better algorithm? +
- Get better features? +++
- Get better data? ++++++

(Leon Bottou)

Scores de correction actuelle en tâche d'extraction d'information

Information type	Accuracy
Entities	90-98%
Attributes	80%
Relations	60-70%
Events	50-60%

(A. McCallum)

Chiffres très optimistes pour des domaines et tâches bien établis (ex. biotechnologie) et des textes non bruités

Connaissances scientifiques et littérature académique

- Le langage spécialisé (terminologie, nomenclature, etc.) est le véhicule principal par lequel la connaissance technique et scientifique est représentée et transmise (Ahmad, 1996)
- Concepts connus vs entités nouvelles vs nomenclatures
- Un terme est l'expression d'un concept scientifique:
  - ➡ Principe Würsterien en terminologie : un concept est exprimé par un terme (pas de synonymie) et un terme ne peut référer qu'à un concept (pas de polysémie).
  - ➡ Pour le TALN et les applications ciblées : l'ambiguité terminologique est constante, la confrontation des concepts, des domaines et des données nourrit le débat scientifique - c'est là que l'on trouve les nouvelles idées et que l'on peut faire émerger de nouvelles hypothèses intéressantes pour les chercheurs

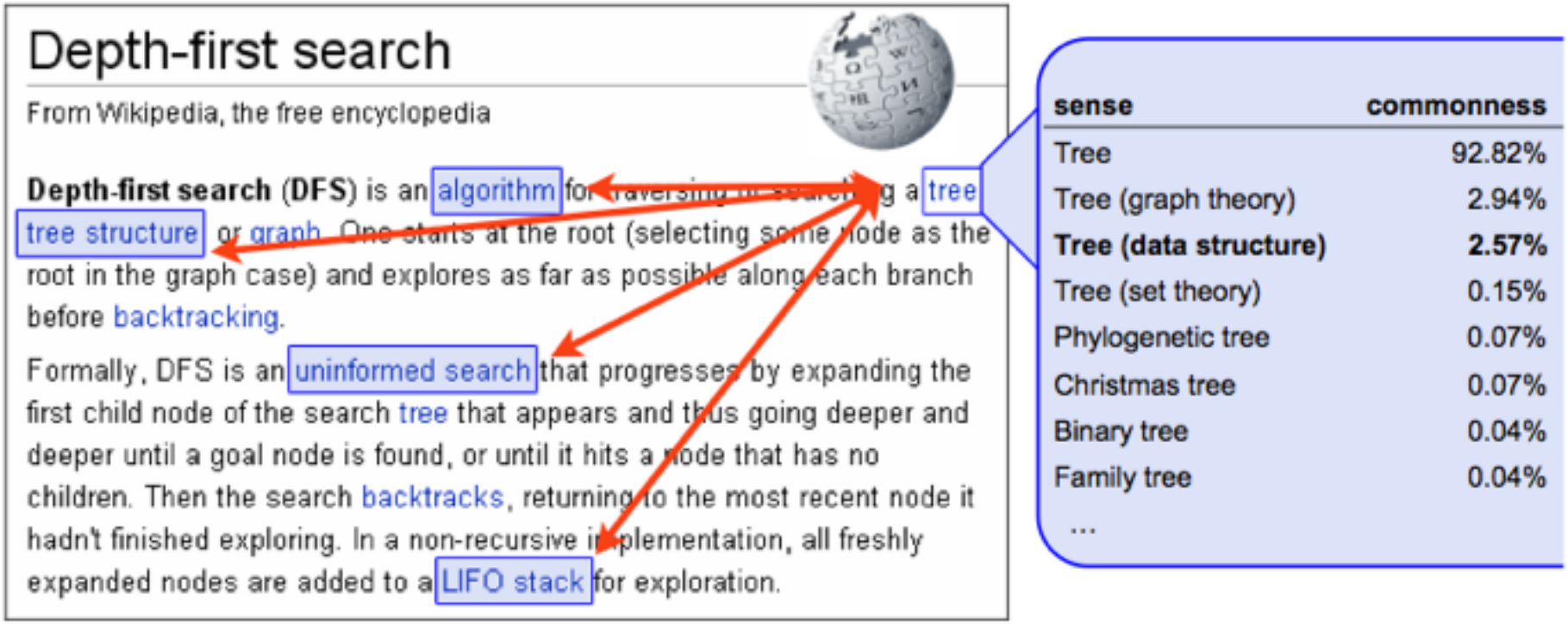
Capturer les concepts connus : (N)ERD

- Identification et résolution des entités avec des bases de connaissances : Wikipedia, FreeBase (Wikidata)
- Couverture immédiate : > 4M de “concepts”, ~150M formes lexicales (pour l’anglais)
- Pas d’autres contraintes d’expertise humaine en ingénierie des connaissances
- Pas de contrainte de domaine
- Multilingue
- Catégorisation suivant les catégories Wikipedia/type Freebase

➡ mais ne couvre que les entités connues!

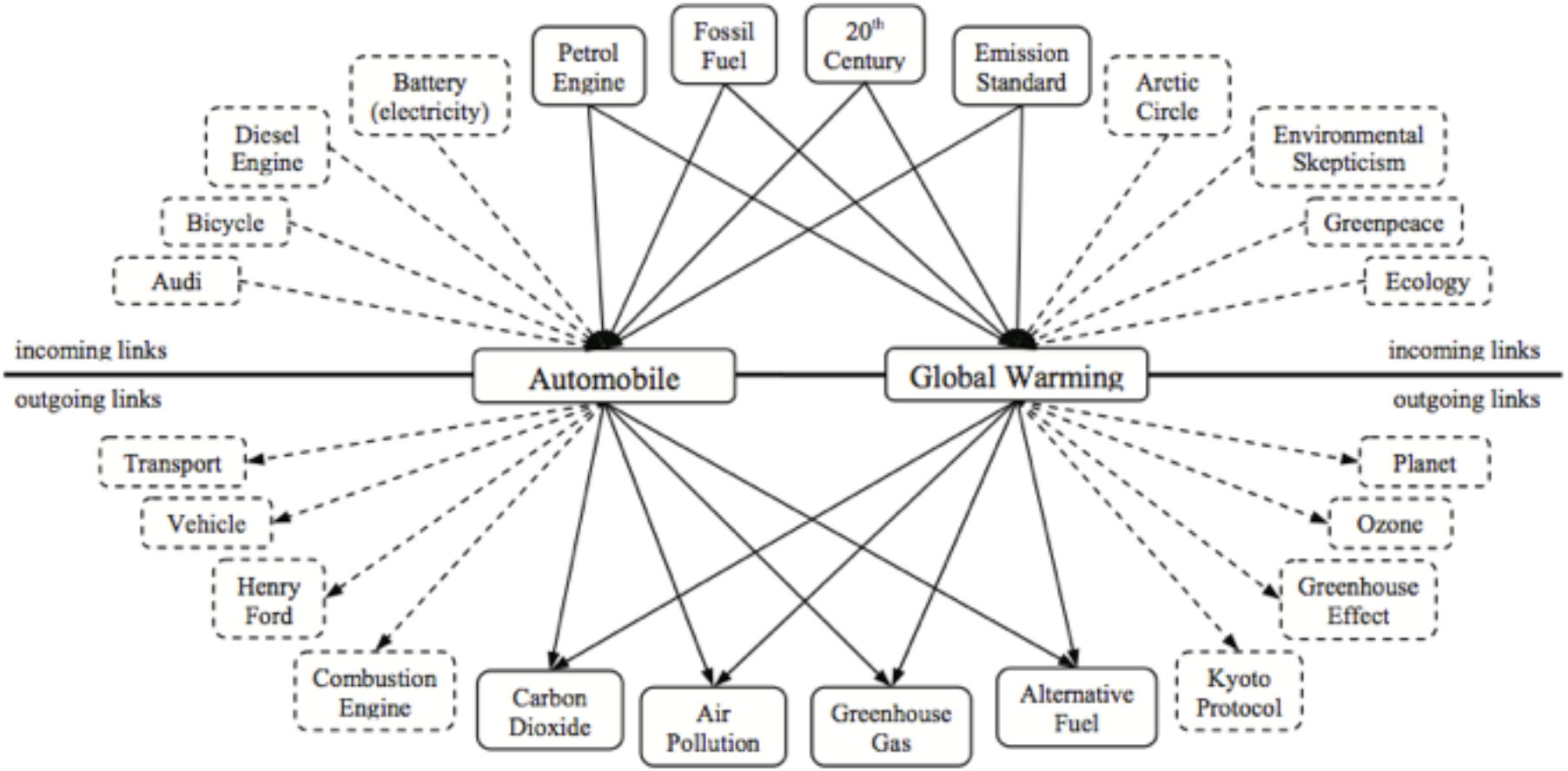
(N)ERD

***Commonness* n'est pas suffisant** : (Milne & Witten, 2008)



NERD

Relatedness: (Milne & Witten, 2009)



Reconnaissance d'entités scientifiques liées à des nomenclatures

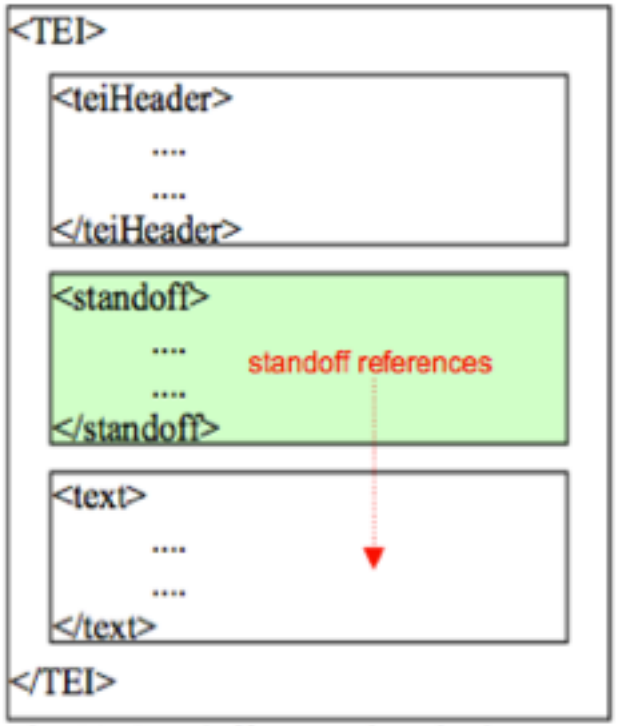
- Nomenclature: un système de nommage permettant de produire des termes et des concepts dans un champ particulier des sciences
- Terminologie et nomenclature = réduire l'impact de la langue naturelle
- Les entités ne peuvent toutes être énumérées par avance car les nomenclatures sont génératives.
  - ➡ PubChem du NIH par exemple reporte plus de 100 millions de formules chimiques dans le corpus brevets mondial
  - ➡ Central en chimie, biologie, astronomie, etc.
  - ➡ Pas de recherche professionnelle d'infos dans ces domaines sans prendre en compte ces termes

Travail du chantier d'usage ISTEX

1. Annotations spécialisées en information scientifique et technique sur un grand échantillon - correspondant à 10 à 20% de l'ensemble du corpus ISTEX
   - (N)ERD : entités connus / catégorisation
   - BIO-GROBID/BEAST : biotech
   - CHEMICAL-GROBID : chimie
   - grobid-quantities : mesures physiques
2. **Encodage TEI stand-off des annotations**
3. Démonstrateur exploitant ces annotations basé sur notre infrastructure anHALytics

Codage TEI d'annotations automatiques

- Persistance et réutilisation des annotations
- Nécessité de schémas détaillés d'annotations, agnostique -> TEI
- Annotations *standoff* en TEI avec offset
- Gestion de multiples annotations concurrentes, cumulables
- Positions préservées

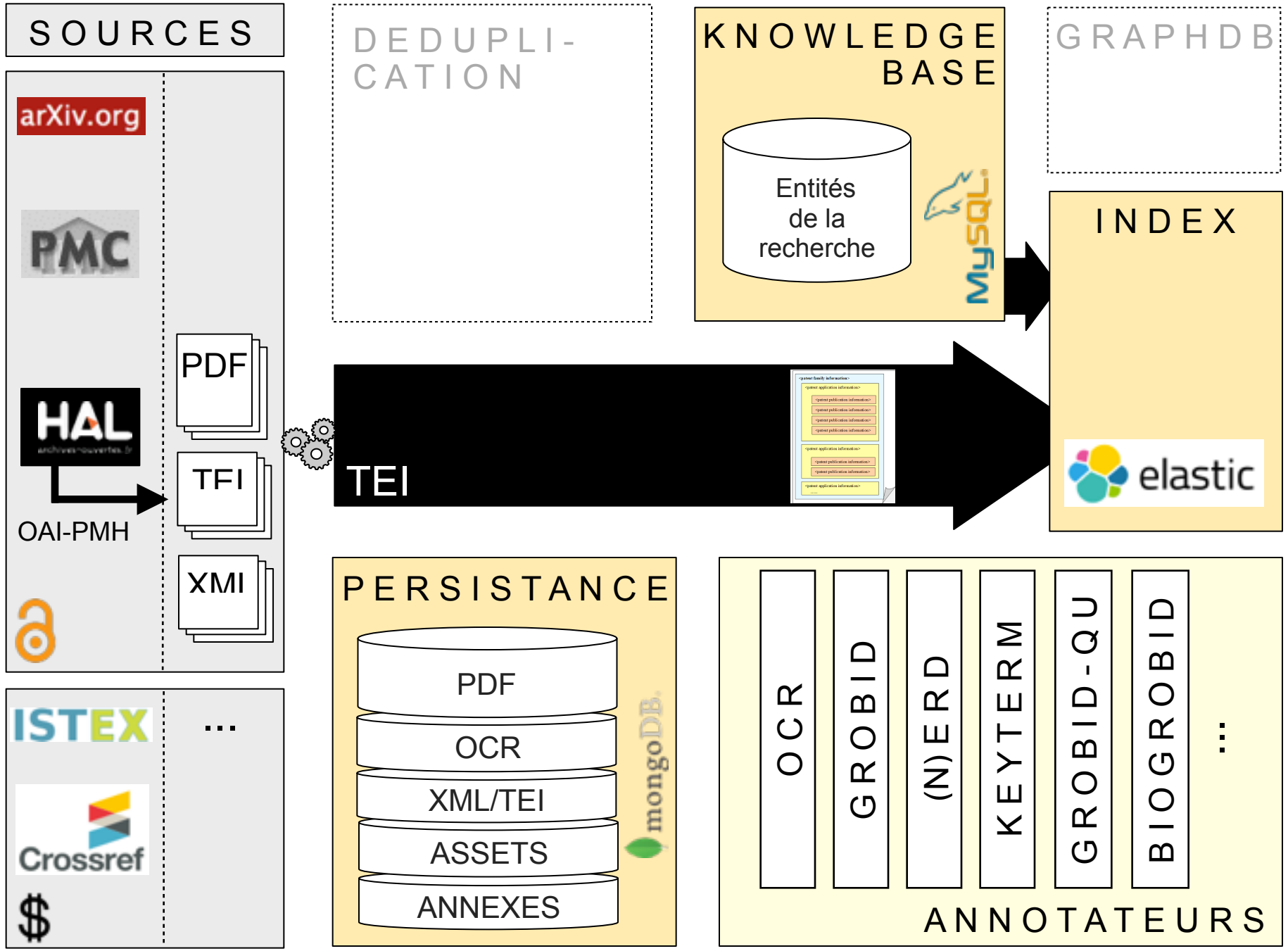


Travail du chantier d'usage ISTEX

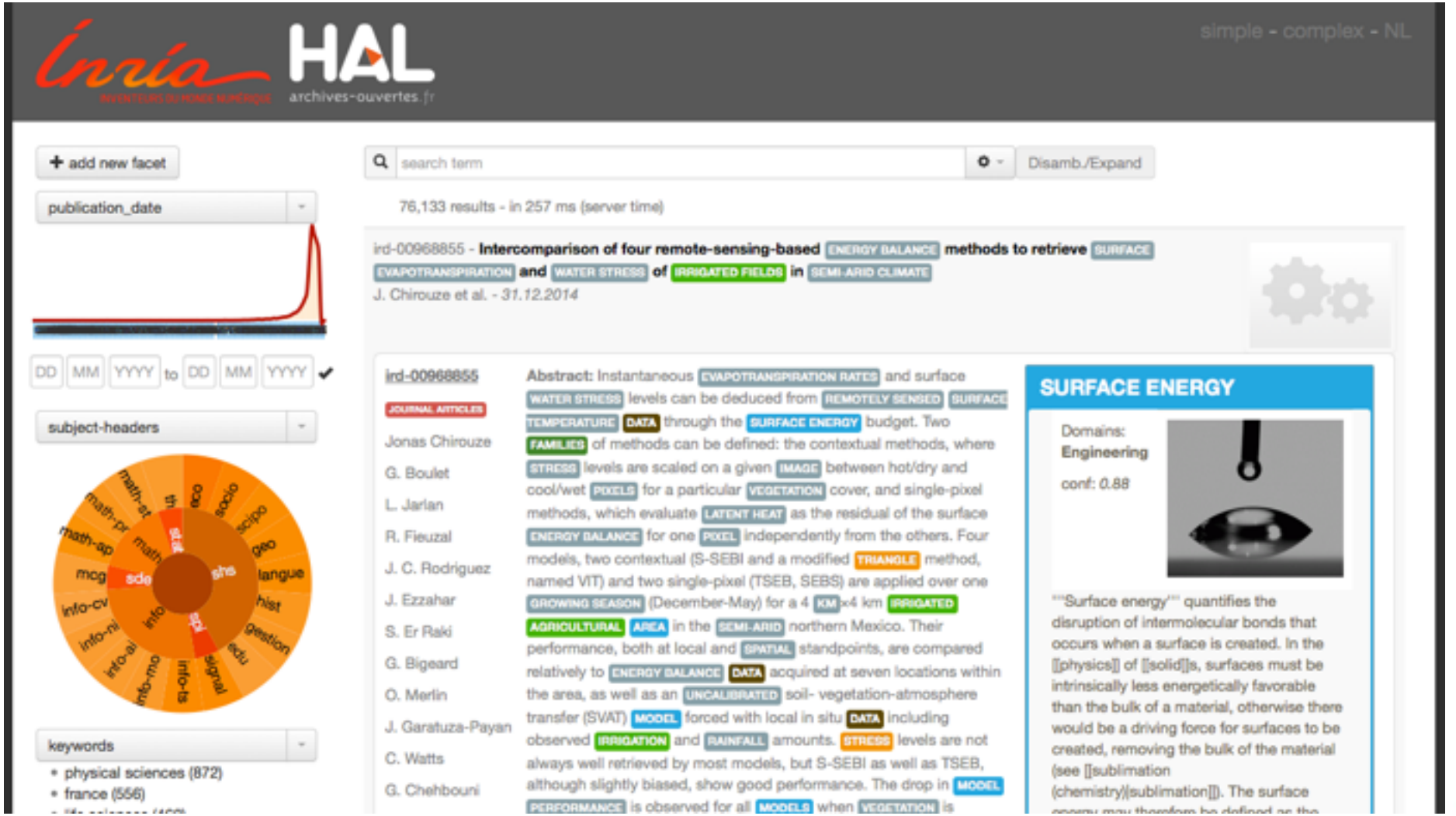
1. Annotations spécialisées en information scientifique et technique sur un grand échantillon - correspondant à 10 à 20% de l'ensemble du corpus ISTEX
   - (N)ERD : entités connus / catégorisation
   - BIO-GROBID/BEAST : biotech
   - CHEMICAL-GROBID : chimie
   - grobid-quantities : mesures physiques
2. Encodage TEI stand-off des annotations
3. **Démonstrateur exploitant ces annotations basé sur notre infrastructure anHALytics**

anHALytics

- Comment améliorer l'engagement des chercheurs français en faveur de l'Accès Libre et HAL?
- Parmi un ensemble d'actions complémentaires, l'idée de l'ADT AnHALytics est d'offrir, en complément du service d'archivage existant, **une plate forme analytique** permettant
  - ➡ la promotion de HAL comme observatoire de l'activité scientifique à différents niveaux de granularité
  - ➡ une motivation supplémentaire pour une politique de mandat de dépôt ambitieuse telle que celle de l'Inria


SOURCES
arXiv.org
PMC
HAL
OAI-PMH
PDF
TFI
XMI
ISTEX
...
Crossref
$
DEDUPLI-
CATION
KNOWLEDGE
BASE
Entités
de la
recherche
MySQL
GRAPHDB
INDEX
elastic
TEI
PERSISTANCE
PDF
OCR
XML/TEI
ASSETS
ANNEXES
mongoDB.
OCR
GROBID
(N)ERD
KEYTERM
GROBID-QU
BIOGROBID
...
ANNOTATEURS

Demo: http://traces1.saclay.inria.fr/anHALytics



Liens

- Grobid: https://github.com/kermitt2/grobid
  - ➡ demo: http://grobid.science-miner.com
- anHALytics: https://github.com/anHALytics
  - ➡ demo: http://traces1.saclay.inria.fr/anHALytics
- (N)ERD: https://github.com/kermitt2/grobid-ner (partial!)
  - ➡ demo: http://nerd.science-miner.com
- GROBID-Quantity: https://github.com/kermitt2/grobid-quantities
  - ➡ demo: http://quantity.science-miner.com
- Keyterm extraction: not yet on GitHub
  - ➡ demo: http://keyterm.science-miner.com
- BEAST demonstrator: not yet on GitHub
  - ➡ demo : en chantier!

Travail du chantier d'usage ISTEX

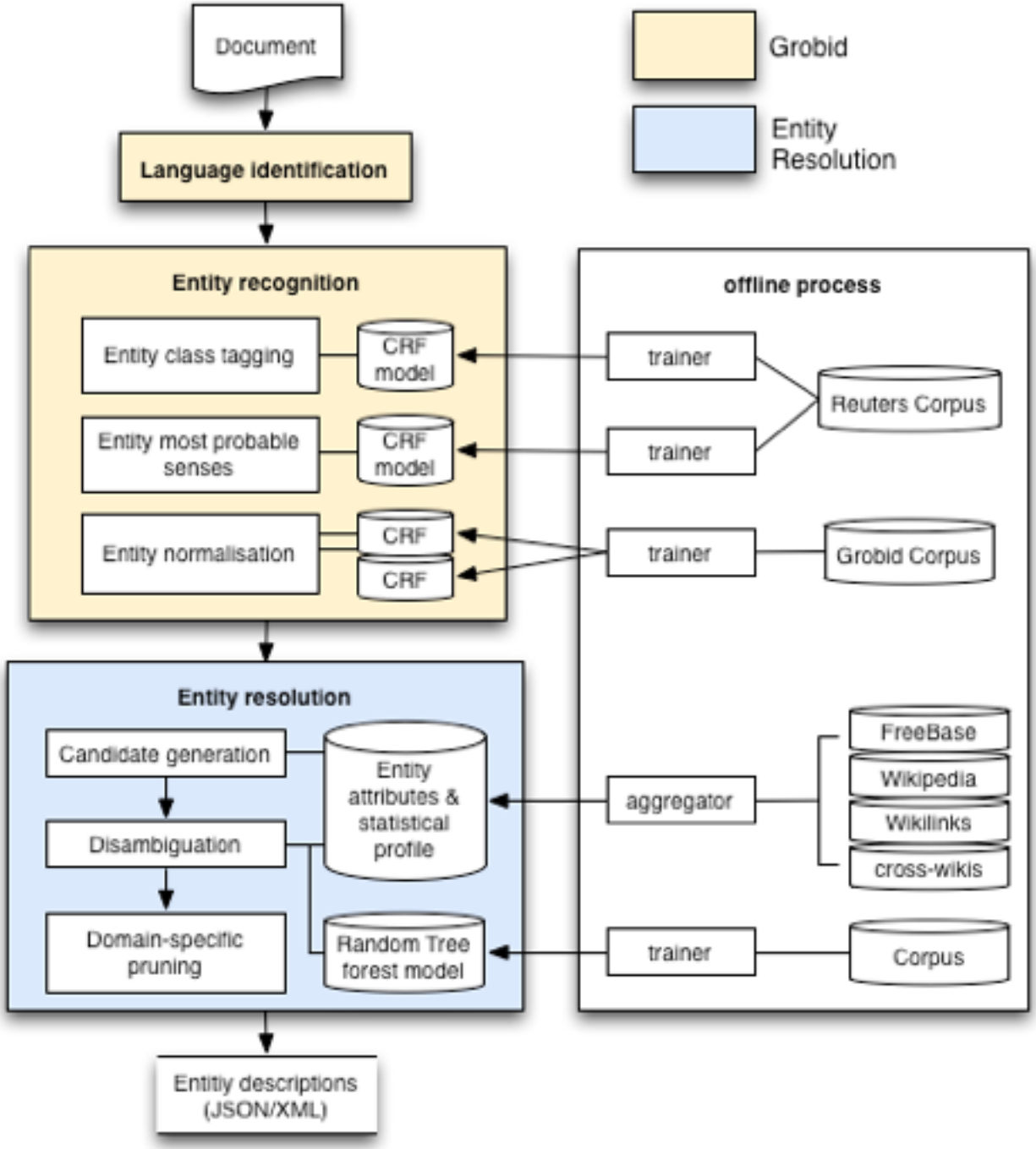
1. Annotations spécialisées en information scientifique et technique sur un grand échantillon - correspondant à 10 à 20% de l'ensemble du corpus ISTEX
   - (N)ERD : entités connus / catégorisation
   - BIO-GROBID/BEAST : biotech
   - CHEMICAL-GROBID : chimie
   - grobid-quantities : mesures physiques
2. Encodage TEI stand-off des annotations
3. Démonstrateur exploitant ces annotations basé sur notre infrastructure anHALytics

Références

- Evans, J., Rzhetsky, A. "*Machine Science*". Science, vol. 329. no. 5990, pp. 399 - 400 (2010).

- Tony Hey, Stewart Tansley, and Kristin Tolle. "*The Fourth Paradigm: Data-Intensive Scientific Discovery*", (2009).

- Younger, S. M. "*Supercomputing and the Human Endeavor: The Coming Scientific Revolution in How We Use Machines to Help Us Think*". White paper, Los Alamos National Laboratory, (2001).


Document
Language identification
Entity recognition
Entity class tagging
CRF model
Entity most probable senses
CRF model
Entity normalisation
CRF
CRF
Entity resolution
Candidate generation
Disambiguation
Domain-specific pruning
Entity attributes & statistical profile
Random Tree forest model
Entitiy descriptions (JSON/XML)
Grobid
Entity Resolution
offline process
trainer
trainer
trainer
trainer
aggregator
trainer
Reuters Corpus
Grobid Corpus
FreeBase
Wikipedia
Wikilinks
cross-wikis
Corpus

NERD

***Features* utilisées pour la désambiguisation générées pour chaque candidat**

- commonness (probabilité du concept pour un terme)
- relatedness with local context
- relatedness with glocal context
- FreeBase type
- Inverse probability (probabilité du terme pour exprimer le concept)
- NER class
- NER sense
- *KL divergence* entre le texte de définition du candidat (par exemple article Wikipedia) et le contexte textuel de mention du terme

➔ Approche ***Machine Learning* supervisée** : Les features sont utilisées pour entrainer un Random Tree Forest à l'aide de données d'entrainement existants (Yahoo! Webscope L24, AIDA/CoNNL-YAGO and ERD-50)

➔ Approche ***Machine Learning* non supervisée** : entrainement complété par une sélection d'articles Wikipedia pour exploiter les ancres/liens vers articles

CENDARI
COLLABORATIVE EUROPEAN DIGITAL ARCHIVE INFRASTRUCTURE

test front-end for ElasticSearch

simple - complex - NL

+ add new facet

normalizedDate

17 18

DD MM YYYY to DD MM YYYY

Library
- Bridgeman Digital Images (6)
- NLS digital collections (2)

SubjectTerms
- German (8)
- fighter pilot (6)
- World War One (6)
- Photograph - 20th century (6)
- Great War (6)
- First World War (6)
- portrait (5)
- male (5)
- Armed Forces - Airforce (5)
- luftwaffe (4)

Language

red baron

Disamb./Expand

8 results - in 16 ms (server time)

The Red Baron (b/w photo).
Anonymous - Published date unknown
*... Photographer false 20th false Bridgeman Digital Images WAR & MILITARY SCENES: 20TH CENTURY UNDETCOP German Photographer Publisher The **Red**...*
*... **Baron** (b/w photo) fighter pilot Armed Forces - Airforce Photograph - 20th century hero uniform aeroplane airplane cap male Air...*

The Red Baron (b/w photo).
Anonymous - Published date unknown
*... false 20th false Bridgeman Digital Images WAR & MILITARY SCENES: 20TH CENTURY UNDETCOP German Photographer Publisher The **Red Baron**...*

Captain Baron VON RICHTHOFEN landing his FOKKER TRIPLANE (b/w photo).
Anonymous - Published date unknown
*... travel Historical Events 1900 - present day World War One German World War I and World War II First World War **Red Baron** UNDETCOP ...*
*... & MILITARY SCENES: 20TH CENTURY UNDETCOP German Photographer Publisher Captain **Baron** von Richthofen landing his Fokker Triplane (b/w photo...*

Author(s): German Photographer

Subject Terms: World War I (1914-18), fighter pilot, Photograph - 20th century, aeroplane, airplane, luftwaffe, Air, fighter plane, Great War, Air travel, Historical Events 1900 - present day, World War One, German, World War I and World War II, First World War, Red Baron
Copyright: undetermined copyright

FOKKER TRIPLANE
Type: **artifact**
Sub-type: **undefined**
conf: *0.856*