Chantier d'usage ISTEX

Inria-Alpage

(équipe ALMAnaCH au 01/01/2017)

Équipe

Achraf Azhar

ISTEX

Patrice Lopez

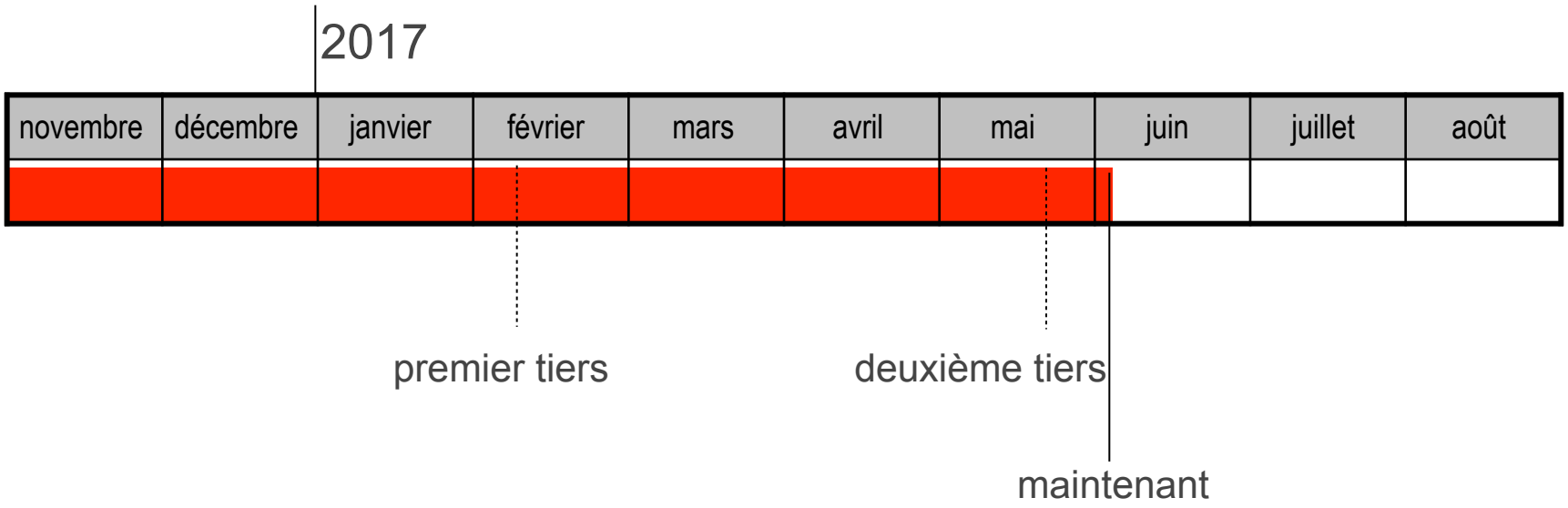
science-miner

Luca Foppiano

Laurent Romary

Avant propos

- Nos travaux pour le chantier d'usage :



Le TDM scientifique aujourd'hui

- Fouille de textes scientifiques : rendre exploitable l'information contenue dans la littérature scientifique et technique
  - ➡ Amélioration des outils de recherche d'information
  - ➡ Construction de bases de connaissances
  - ➡ Production automatique d'hypothèses scientifiques

Le TDM scientifique aujourd'hui

- Fouille de textes scientifiques : rendre exploitable l'information contenue dans la littérature scientifique et technique
  - ➡ Amélioration des outils de recherche d'information
  - ➡ Construction de bases de connaissances
  - ➡ Production automatique d'hypothèses scientifiques

- Des résultats bien concrets :
  - ➡ Outils de recherche documentaire
  - ➡ Services bibliographiques/documentaires
  - ➡ Découverte et développement de médicaments (facteur 4)

Le TDM scientifique aujourd'hui

- Fouille de textes scientifiques : rendre exploitable l'information contenue dans la littérature scientifique et technique
  - ➡ Amélioration des outils de recherche d'information
  - ➡ Construction de bases de connaissances
  - ➡ Production automatique d'hypothèses scientifiques

- Des résultats bien concrets :
  - ➡ Outils de recherche documentaire
  - ➡ Services bibliographiques/documentaires
  - ➡ Découverte et développement de médicaments (facteur 4)

- Mais très couteux :
  - ➡ Expertises pour la construction de règles/patterns, bases de connaissances structurées exploitables à grandes échelles

Nos travaux en TDM

- Structuration automatique de documents brutes (PDF) : GROBID (ResearchGate, CERN, NASA, HAL, Mendeley, etc.)

Nos travaux en TDM

- Structuration automatique de documents brutes (PDF) : GROBID (ResearchGate, CERN, NASA, HAL, Mendeley, etc.)
- Extraction de termes, concepts et catégories clefs : KeyTerm

Nos travaux en TDM

- Structuration automatique de documents brutes (PDF) : GROBID (ResearchGate, CERN, NASA, HAL, Mendeley, etc.)
- Extraction de termes, concepts et catégories clefs : KeyTerm
- Désambiguïsation d'entités connues : (N)ERD (Wikipedia/data)

Nos travaux en TDM

- Structuration automatique de documents brutes (PDF) : GROBID (ResearchGate, CERN, NASA, HAL, Mendeley, etc.)
- Extraction de termes, concepts et catégories clefs : KeyTerm
- Désambiguïsation d'entités connues : (N)ERD (Wikipedia/data)
- Ingestion documentaire TEI - Pub2TEI

Nos travaux en TDM

- Structuration automatique de documents brutes (PDF) : GROBID (ResearchGate, CERN, NASA, HAL, Mendeley, etc.)
- Extraction de termes, concepts et catégories clefs : KeyTerm
- Désambiguïsation d'entités connues : (N)ERD (Wikipedia/data)
- Ingestion documentaire TEI - Pub2TEI
- Exploitation de HAL comme corpus pour le TDM et l'observation de l'activité scientifique française (anHALytics)

Nos travaux en TDM

- Structuration automatique de documents brutes (PDF) : GROBID (ResearchGate, CERN, NASA, HAL, Mendeley, etc.)
- Extraction de termes, concepts et catégories clefs : KeyTerm
- Désambiguïsation d'entités connues : (N)ERD (Wikipedia/data)
- Ingestion documentaire TEI - Pub2TEI
- Exploitation de HAL comme corpus pour le TDM et l'observation de l'activité scientifique française (anHALytics)
- Extractions spécialisées/nomenclaturées : extraction d'entités biotech, astronomiques, chimiques, de mesures scientifiques

Nos travaux en TDM

- Structuration automatique de documents brutes (PDF) : GROBID (ResearchGate, CERN, NASA, HAL, Mendeley, etc.)
- Extraction de termes, concepts et catégories clefs : KeyTerm
- Désambiguïsation d'entités connues : (N)ERD (Wikipedia/data)
- Ingestion documentaire TEI - Pub2TEI
- Exploitation de HAL comme corpus pour le TDM et l'observation de l'activité scientifique française (anHALytics)
- Extractions spécialisées/nomenclaturées : extraction d'entités biotech, astronomiques, chimiques, de mesures scientifiques
- Déduplication (*matching*) d'entités scientifiques

Objectifs du chantier d'usage ISTEX

- Utilité du corpus ISTEX pour la conception, l'industrialisation et la montée en charge d'annotateurs
  - ➡ annotations sur un grand échantillon - correspondant à 10 à 20% de l'ensemble du corpus ISTEX
  - ➡ croiser les disciplines,
  - ➡ utiliser des sources réelles (PDF éditeurs)
- Encodage TEI stand-off des annotations
- Démonstrateur exploitant ces annotations basé sur notre infrastructure anHALytics

Priorisation

- L'équipe ISTEX R&D a déjà entrepris depuis 2014 le ré-entrainement et l'intégration de GROBID dans ISTEX
  - ➡ 1 million de PDF traités en 24h (9 threads) - 11,5 PDF/s
- L'équipe ISTEX R&D expérimente depuis début 2016 les modules d'extraction de termes, concepts et catégories clefs (grobid-keyterm) et de reconnaissance et extraction d'entités (N)ERD
- Par complémentarité, pour ce chantier d'usage, nous nous concentrons sur l'annotateurs grobid-quantities et son expérimentation sur le corpus ISTEX

Identification et normalisation des mesures physiques

- Origines :
  - ➡ besoin identifié pour les brevets
  - ➡ demande du JPL (NASA), utilisateur de GROBID
- Besoin transversal à toutes les disciplines scientifiques
- Solutions existantes limitées, propriétaires, par règles manuelles

démo

http://quantity.science-miner.com

Identification et normalisation des mesures physiques

- Extension grobid-quantities
- Notre approche :
  - ➡ machine learning, cascades LC-CRF
  - ➡ reconnaissance valeurs, intervalles, listes
  - ➡ couverture d'un maximum de types de mesures
  - ➡ normalisation en unités SI (JSR-363, ISO 80000, ...)
  - ➡ annotation de texte, XML mais aussi PDF
  - ➡ open source Apache 2
- 1000 mots par seconde (1 thread)
- Reconnaissance des substances quantifiées

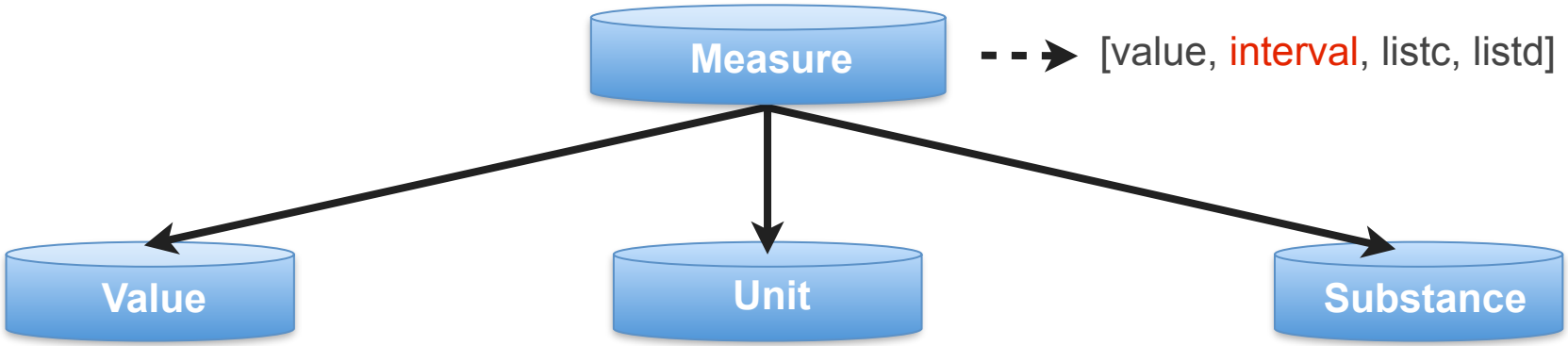
démo

http://quantity.science-miner.com

JPL NASA integration https://github.com/khundman/marve

Modèles CRF dans grobid-quantities

... and 200–700 µmol/l of sodium valproate with ...



200
700

prefix: µ
unit: mol
unit: liter

sodium valproate

type: amount of substance

normalization: [0.2, 0.7]
normalized unit (SI): mol/m3
normalized: sodium valproate (W 650844)

from doi:10.1186/1475-2832-3-13

Subjects and Study Design

All subjects gave full informed consent, and both studies were approved by the ethics committee at the University of Alberta. Healthy controls were examined using a detailed, but non-standardized, psychiatric interview. They were excluded if there was any personal history, or immediate family history, of psychiatric disorder. For patients, diagnoses were made using DSM-IV criteria for Bipolar Disorder Type I or Type II following detailed psychiatric interview, with additional information being available in almost all cases from long-term psychiatric clinic records. They also had to be taking a dose of either lithium or valproate which maintained their blood levels within the ranges of 0.4–1.2 mmol/l for lithium and 200–700 μmol/l for sodium valproate. Serum lithium and valproate levels were also measured on the day of MRS scanning. Other medications taken by the patient were noted. In the second part of the study the same criteria were used, except that only patients meeting diagnostic criteria for Bipolar Disorder Type I were included, and they had to be on sodium valproate monotherapy. This was done to examine Bipolar Type I patients in more detail, and to remove a possible confounding variable. All patients had to be euthymic for the previous 3 months, as determined

Interval

quantity type: **concentration**

raw: from **200** to **700**

raw unit name: **μmol/l**

normalized: from **0.2** to **0.7**

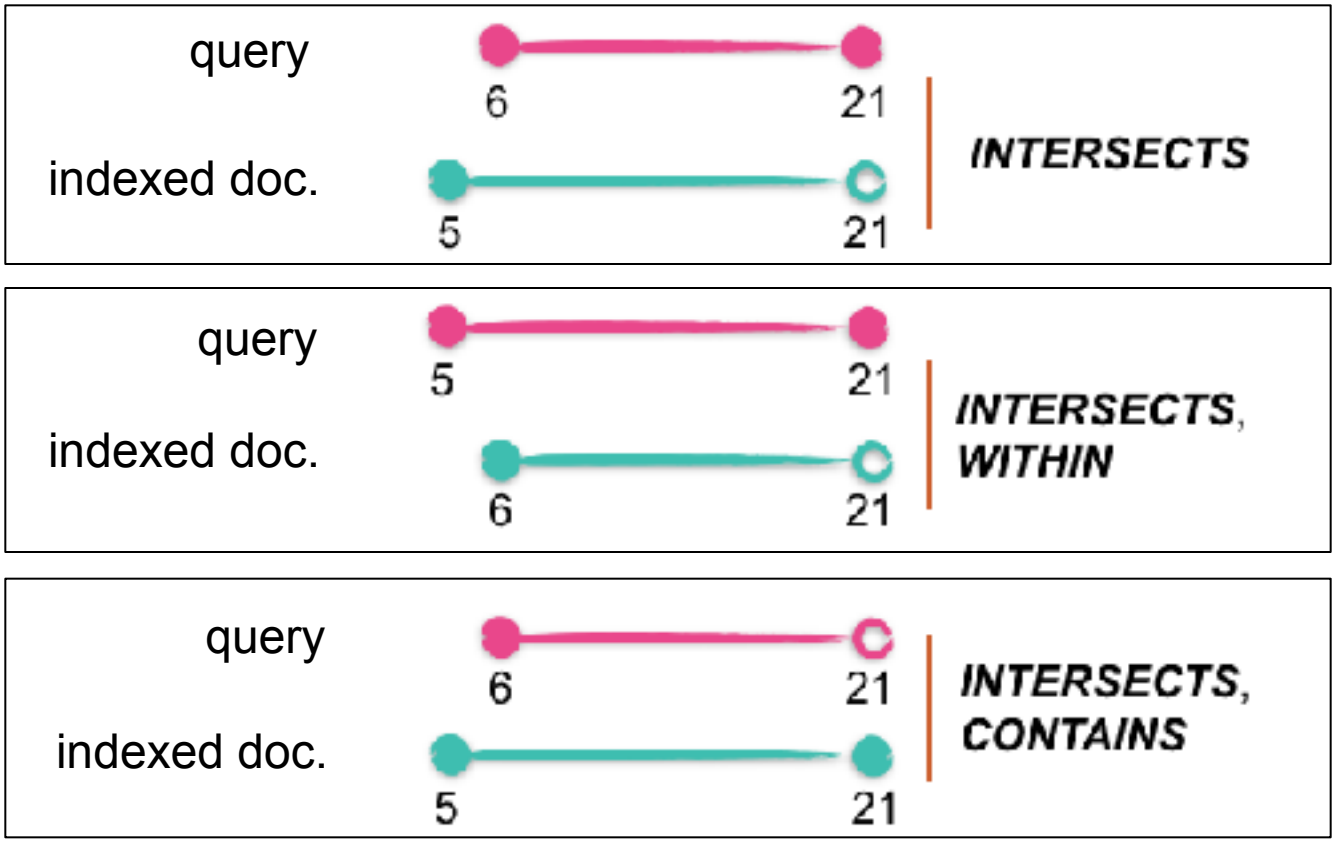
normalized unit: **mol/m³**

quantified (experimental):

Recherche documentaire par quantités

- Lucene : recherche de valeurs discrètes dans des intervalles
- Uwe Schindler & *BKD tree structures* dans Lucene :
  - ➡ possibilité de cherche spatiale pour objets 2D, voir 3D
  - ➡ fonctionne aussi en 1D : recherche d'intervalles par intervalles et *scoring* spécifique

Recherche documentaire par quantités



Recherche documentaire par quantités

- Lucene : recherche de valeurs discrètes dans des intervalles
- Uwe Schindler & *BKD tree structures* dans Lucene :
  - ➡ possibilité de cherche spatiale pour objets 2D, voir 3D
  - ➡ fonctionne aussi en 1D : recherche d'intervalles par intervalles et *scoring* spécifique

Recherche documentaire par quantités

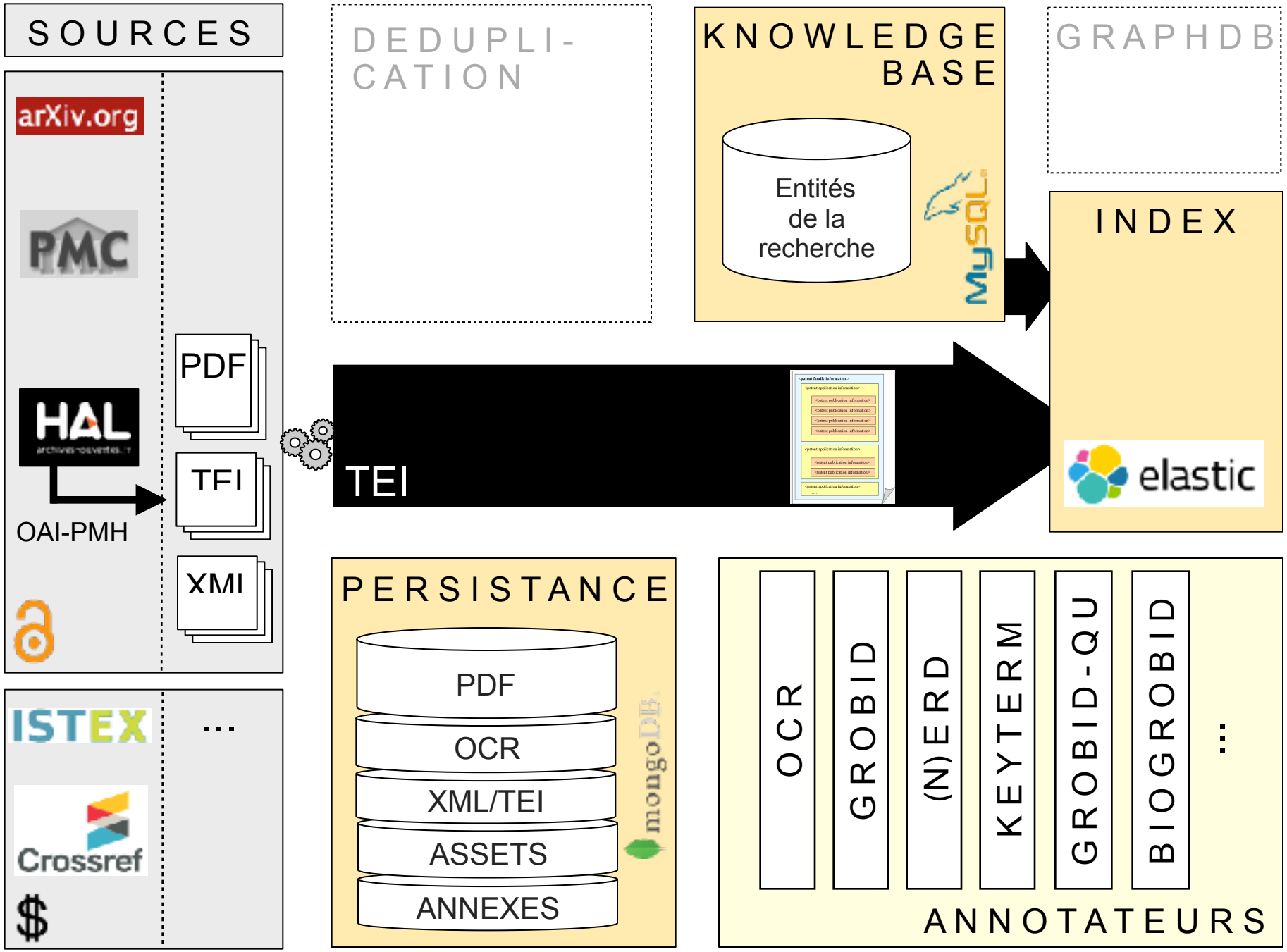
- Lucene : recherche de valeurs discrètes dans des intervalles
- Uwe Schindler & *BKD tree structures* dans Lucene :
  - ➡ possibilité de cherche spatiale pour objets 2D, voir 3D
  - ➡ fonctionne aussi en 1D : recherche d'intervalles par intervalles et *scoring* spécifique
  - ➡ des évolutions très récentes : Lucene 04/2016, Solr 08/2016, ElasticSearch 01/2017

Recherche documentaire par quantités

- Lucene : recherche de valeurs discrètes dans des intervalles
- Uwe Schindler & *BKD tree structures* dans Lucene :
  - ➡ possibilité de cherche spatiale pour objets 2D, voir 3D
  - ➡ fonctionne aussi en 1D : recherche d'intervalles par intervalles et *scoring* spécifique
  - ➡ des évolutions très récentes : Lucene 04/2016, Solr 08/2016, ElasticSearch 01/2017
- Nouvelles perspectives en recherche d'information scientifiques et en text mining
- Exemple : recherche de documents par critères de positionnement 3D d'entités astronomiques dans l'univers

Objectifs du chantier d'usage ISTEX

- Utilité du corpus ISTEX pour la conception, l'industrialisation et la montée en charge d'annotateurs
  - ➡ annotations sur un grand échantillon - correspondant à 10 à 20% de l'ensemble du corpus ISTEX
  - ➡ croiser les disciplines,
  - ➡ utiliser des sources réelles (PDF éditeurs)
- Encodage TEI stand-off des annotations
- Démonstrateur exploitant ces annotations basé sur notre infrastructure anHALytics


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

Ingestion de ressources ISTEX

- Utilisation des catégories WoS et ScienceMetrix pour la sélections de domaines riches en mentions de quantités
- Test et travaux actuels sur un échantillon
- Difficultés :
  - ➡ limitations de l'approche par ZIP : volume, efficacité, stabilité, évolution
  - ➡ suggestions : passer à OAI-PMH avec la définition d'ensembles et sous-ensembles pertinents
  - ➡ suggestions : séparer l'harvesting des métadonnées du téléchargement des PDF

démo

questions?

questions?

questions?

questions?

questions?

questions?

questions?

Le TDM : des opportunités de nouvelles activités pour les documentalistes et bibliothécaires

- Imaginer des nouveaux workflows et de nouvelles applications intégrant ces techniques
- Curation de données de la recherche
- Curation de données d'apprentissage pour les outils d'annotations
- Tester et évaluer ces outils
  - ➡ test indépendant, benchmarking
  - ➡ test end-to-end
  - ➡ test A/B et méthodologies

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

Approche

- Les approches par apprentissage automatique 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)

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

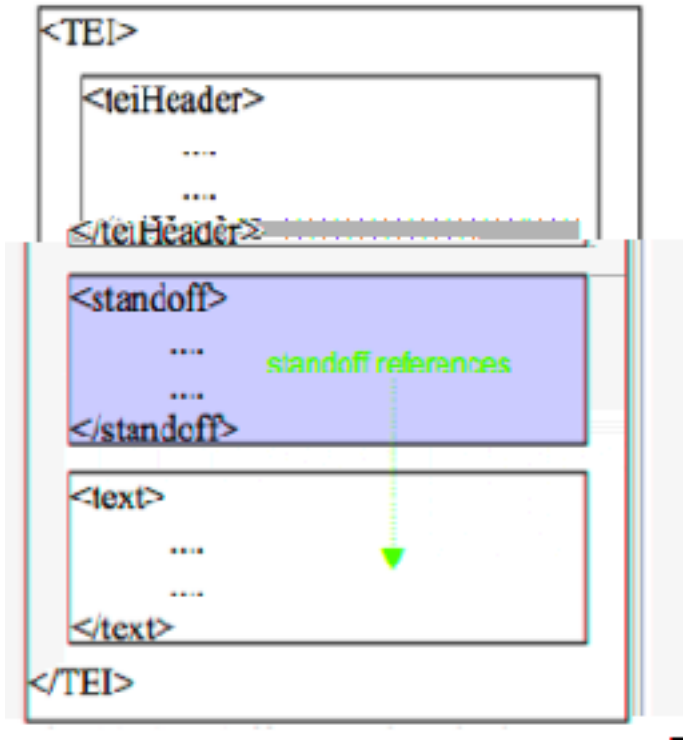
➡ mais ne couvre que des entités connues!

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

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**

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!