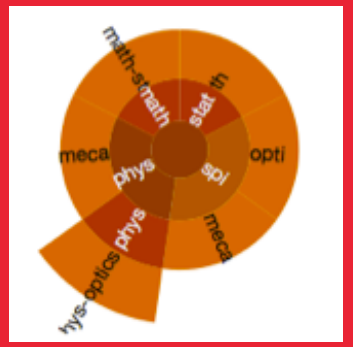



# AnHALytics : plateforme analytique pour les archives ouvertes

## Journée Archives ouvertes et bases de publications

# Motivation

- 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,

# Exemples de services visés

- Rapports d'activité et de collaborations multi-niveaux - institut, laboratoire, équipe, individu
- Vues agrégées thématiques et multi-critères
- Cartographies technologiques et scientifiques
- Évolution temporelle des thématiques de recherche
- Intersection thématique et opportunité de collaboration interdisciplinaire

# 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)
- Difficulté d'exploitation du format PDF / pauvreté et incohérence des metadonnées

GROBID

- "*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

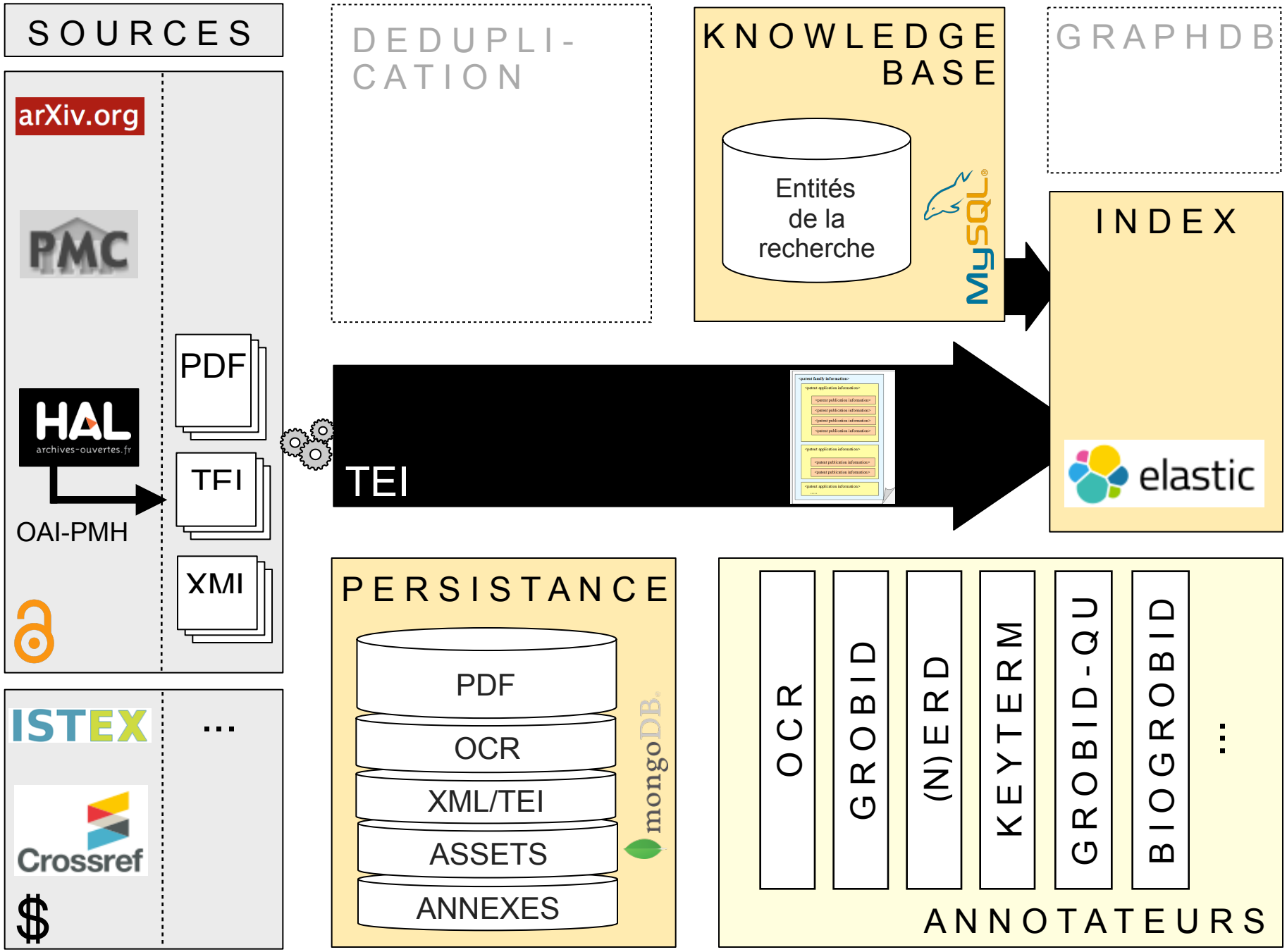
- Gestion de la volumétrie : million de documents, milliards d'annotations, graphe avec milliard de noeuds
- Gestion de l'incertitude et du bruit: les meilleures techniques de fouille de textes font des erreurs, beaucoup d'erreurs...
- *Machine learning* : domaines et scénarios plus complexes et ouverts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

# Tirer profit du jeu de données

- 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

# Outils et techniques

- Nos résultats de l'état de l'art en text mining :
  - ➡ GROBID, référence en analyse d'articles PDF,
  - ➡ NERD , désambiguïsation automatique d'entités scientifiques
  - ➡ KeyTerm, extraction de termes clefs (premier à SemEval 2010)
  - ➡ ….
- Une infrastructure documentaire performante : Moissonage/ Stockage/Indexation journalier de l'intégralité de HAL – fouille de données sur les dépôts avec fulltexts (~350k)
- Une base de connaissances pour les entités de la recherche:
  - ➡ Exploitation des metadonnées HAL,
  - ➡ Enrichissement par extraction automatique et par croisement avec des bases de références IST


SOURCES
arXiv.org
PMC
HAL
archives-ouvertes.fr
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

# Demo

# Perspective

- Améliorer les interfaces d'anHALytics :
  - ➡ Tableaux de bord par entité/domaine de recherche.,
  - ➡ Recherche pointue ajustable.
  - ➡ Plus de facettes de recherche
  - ➡ ….
- Moteur de recherche de figures
- Exploitation des mot-clés lecteurs
- Construire un web de données
- Désambiguisation des auteurs et des structures de recherche pour le référentiel HAL
- Test de la montée en charge de la plateforme à travers le chantier d'usage ISTEX
- Collaborer avec d'autres projets

# Conclusion

- Des outils état de l'art, OCR/apprentissage/big data/NLP
- Open source,
- Un plateforme analytique temps-réel

➡ Pour promouvoir et encourager le dépôt et le partage des publications, participer au rayonnement des travaux des chercheurs en utilisant des indicateurs, faire des recommandations et faciliter la prise de decision.

# 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