Visa TM

Fredoc 2018

Bases de ressources numériques et services auxchercheurs. Avec ISTEX et OpenMintedD, l'alliance pour une infrastructure de text-mining

Laurence el khouri – DIST/CNRS

Stéphane Schneider - INIST-CNRS

ISTEX
L'excellence documentaire pour tous
ANR-10-IDEX-0004-02

LE TDM et ses enjeux

Visa TM

L'accès à la production scientifique
Un environnement en mutation

ISTEX
L'excellence documentaire pour tous
ANR-10-IDEX-0004-02

- **La transformation numérique : une urgence**

Evolution des objets accessibles

Evolution des méthodes d'accès

- **Massification des informations**

Exploration des contenus :

Extraction d'information et de connaissance

Outils de visualisation : information précise, tendances …

- **Méthodes et traitements informatiques pour analyser le sens de textes** en langage naturel pour en donner une représentation utilisable par les humains et les ordinateurs. Spécialisation de la fouille de données (data mining) qui fait appel aux techniques de l'Intelligence Artificielle, du Traitement Automatique des Langues et des Statistiques.

Visa TM

Contexte science ouverte

ISTEX
L'excellence documentaire pour tous
ANR-10-IDEX-0004-02

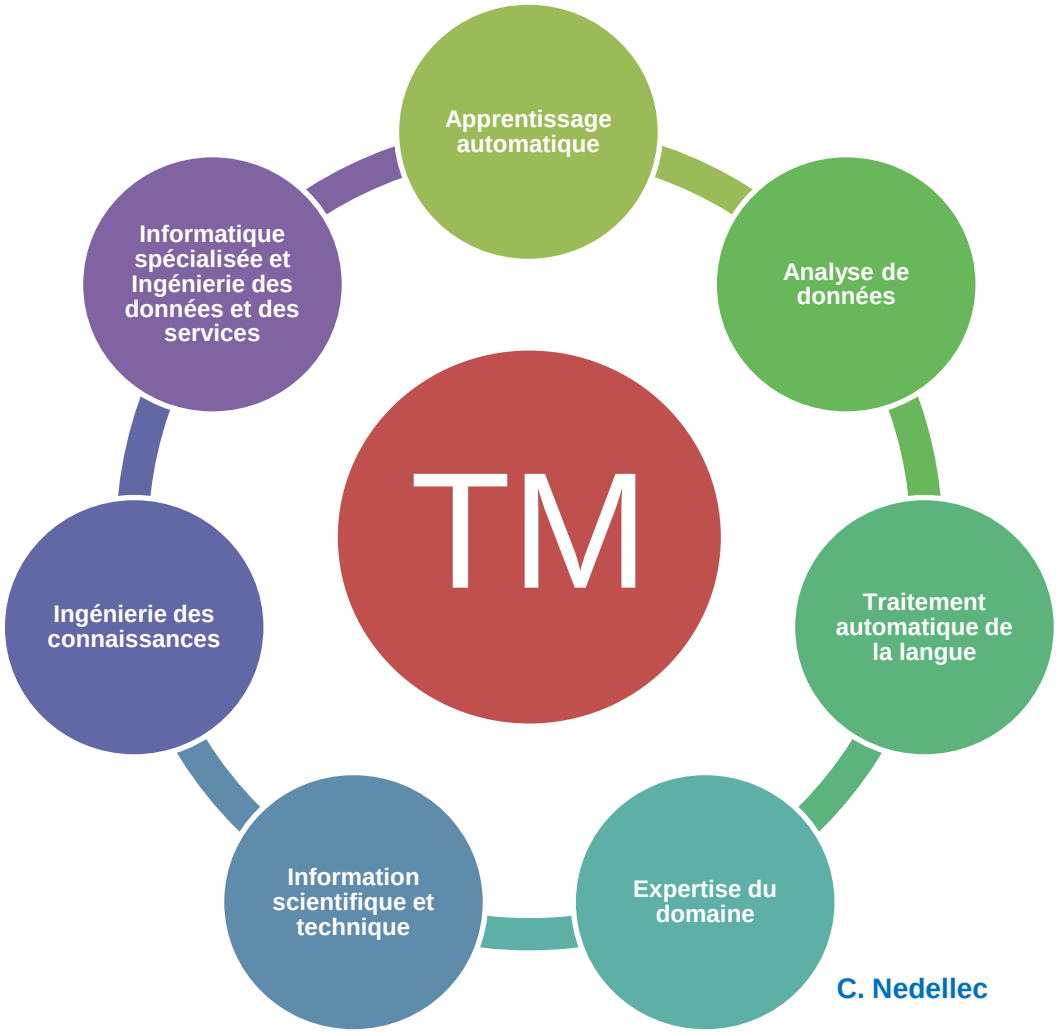
- **Article 38, loi pour une république numérique**
- **Article 3, révision de la directive DADVSI**

« The right to read is the right to mine »

- **Guide d'application de la loi :**
  - analyse des besoins
  - recommandations aux chercheurs

TM au confluent de nombreux domaines

ANR-10-IDEX-0004-02



C. Nedellec

ANR-10-IDEX-0004-02

Objectifs initiaux

ISTEX
L'excellence documentaire pour tous
ANR-10-IDEX-0004-02

Création d'une plateforme nationale destinée à l'ensemble de la communauté de l'ESR intégrant :

- **Des acquisitions pérennes,** sous forme de licence nationale, de ressources documentaires multidisciplinaires **autorisant l'exploration de texte**
- **Une cohérence systémique de l'ensemble des droits** sur les ressources ISTEX et sur les ressources courantes
- **L'agrégation des ressources** au sein d'une plateforme nationale apportant une plus-value basée sur le traitement des données en texte intégral
- **Une plateforme interopérable** avec celles des établissements et organismes du paysage français de l'ESR
- **L'offre de services et usages complémentaires :** traitement des données , extraction de données, fouille de textes, production de synthèses documentaires et de corpus terminologiques …

abes agence bibliographique de l'enseignement supérieur — Négociation, acquisition des ressources, signalement, gestion des accès et des droits

couperin.org Consortium Unitié des Etablissements Universitaires et de Recherche pour l Accès aux Publications Numériques — Recueil et analyse des besoins, lancement des appels à propositions, évaluation des offres et ressources, pré-sélection, détermination prix-cibles, support aux négociations

cnrs — Pilotage du projet (DIST), Développement de la plateforme (INIST)

Université de Lorraine, Conférence des présidents d'université — Coordination des services à valeur ajoutée et chantiers d'usage

Visa TM

ISTEX fermé ?

ISTEX
L'excellence documentaire pour tous
ANR-10-IDEX-0004-02

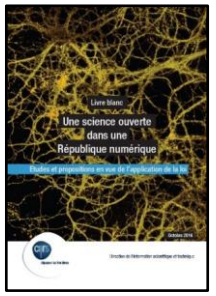
- ✓ **ESR français**
- ✓ **Contenus fermés – copyright cédé aux éditeurs**
- ✓ **Appropriation des contenus sans limite de temps**
- ✓ **Chercheurs/auteurs choisissez votre licence : CC0**

Copyright partagé :

Copyright cédé :

- ✓ **Connexion à des bases ouvertes**
- ✓ **Permis rédaction article 38**

Livre blanc 1 - mars 2016
Livre blanc 2 – oct 2016



- ✓ **Contenu certifié fouillable**

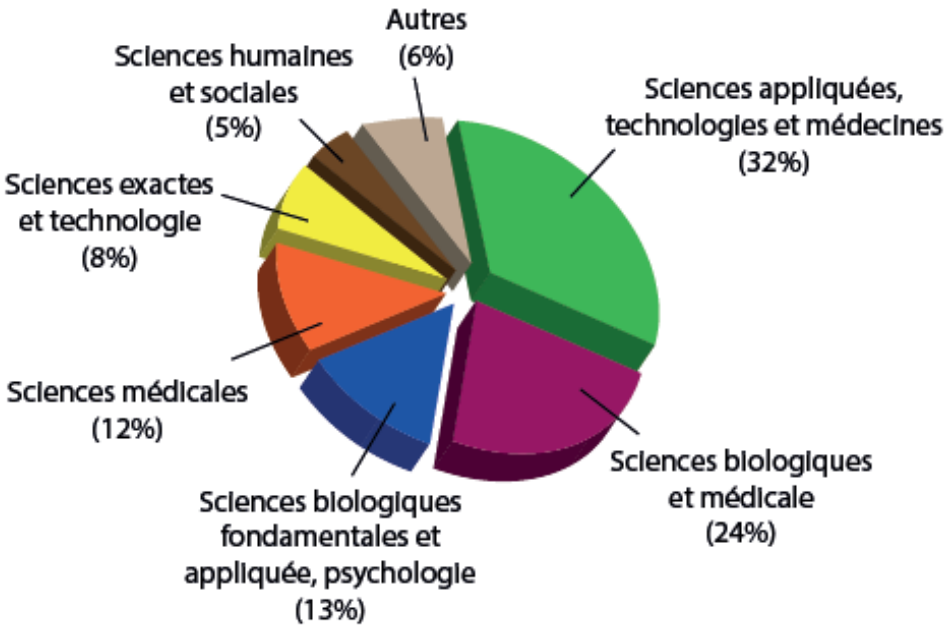
Bac à sable pour text mining

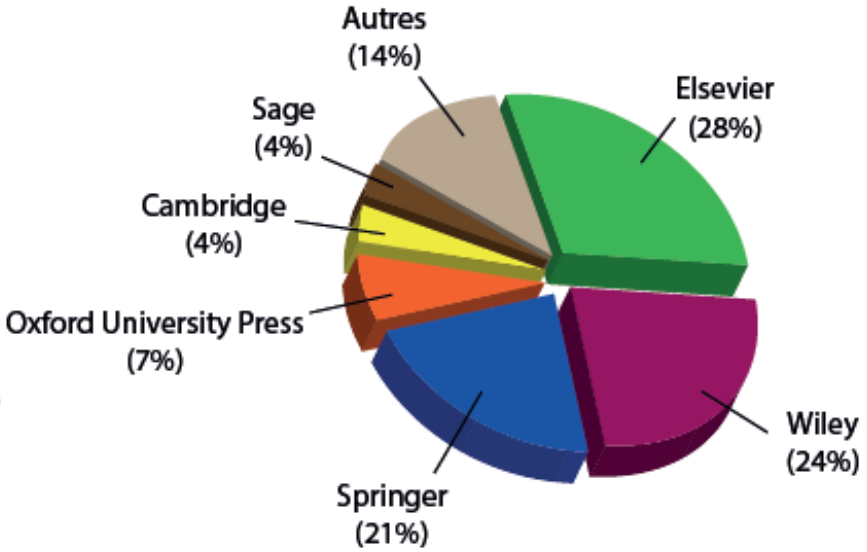
Visa TM

Ressources documentaires

ISTEX
L'excellence documentaire pour tous
ANR-10-IDEX-0004-02

21,6 Millions d'objets disponibles	21 Corpus différents





> 9000 revues	~345 000 monographies*
	dont >300 000 sur ECCO/EEBO

Visa TM

Une plateforme innovante 1/2

ISTEX
L'excellence documentaire pour tous
ANR-10-IDEX-0004-02

- **Widget, plug-in et API** pour un usage nomade et local selon les besoins
- **corpus multidisciplinaire, multilingue** homogénéisé, enrichi et fouillable (indépendance aux abonnements en cours)
  - **Complémentaire** avec les abonnements courants et la production en open access :
- Accès systématique vers le **texte intégral** du document
- Exposition dans **le Web Sémantique**
- **Moteur de recherche** adapté aux besoins offrant des facilités d'interrogation et de téléchargement

Visa TM

Une plateforme innovante 2/2

ISTEX
L'excellence documentaire pour tous
ANR-10-IDEX-0004-02

- Services permettant le **traitement des données** : extraction de données, fouille de textes, production de synthèses documentaires et de corpus terminologiques…
- Stimule le **développement collaboratif d'outils logiciels de TDM** en open source
- **Co-construit et diffuse des ressources linguistiques** : terminologies, ontologies, grammaires
- **Offre des stratégies d'accès aux connaissances scientifiques** par des approches navigationnelles et cartographiques
- **Développe de nouveaux usages et de nouveaux champs de recherches** liés à la numérisation et à la fouille de texte et de données (TM)
- Forme de nouvelles compétences en IST

Visa TM

Démo/Ressources documentaires

ISTEX
L'excellence documentaire pour tous
ANR-10-IDEX-0004-02

Requête : [Démonstrateur]

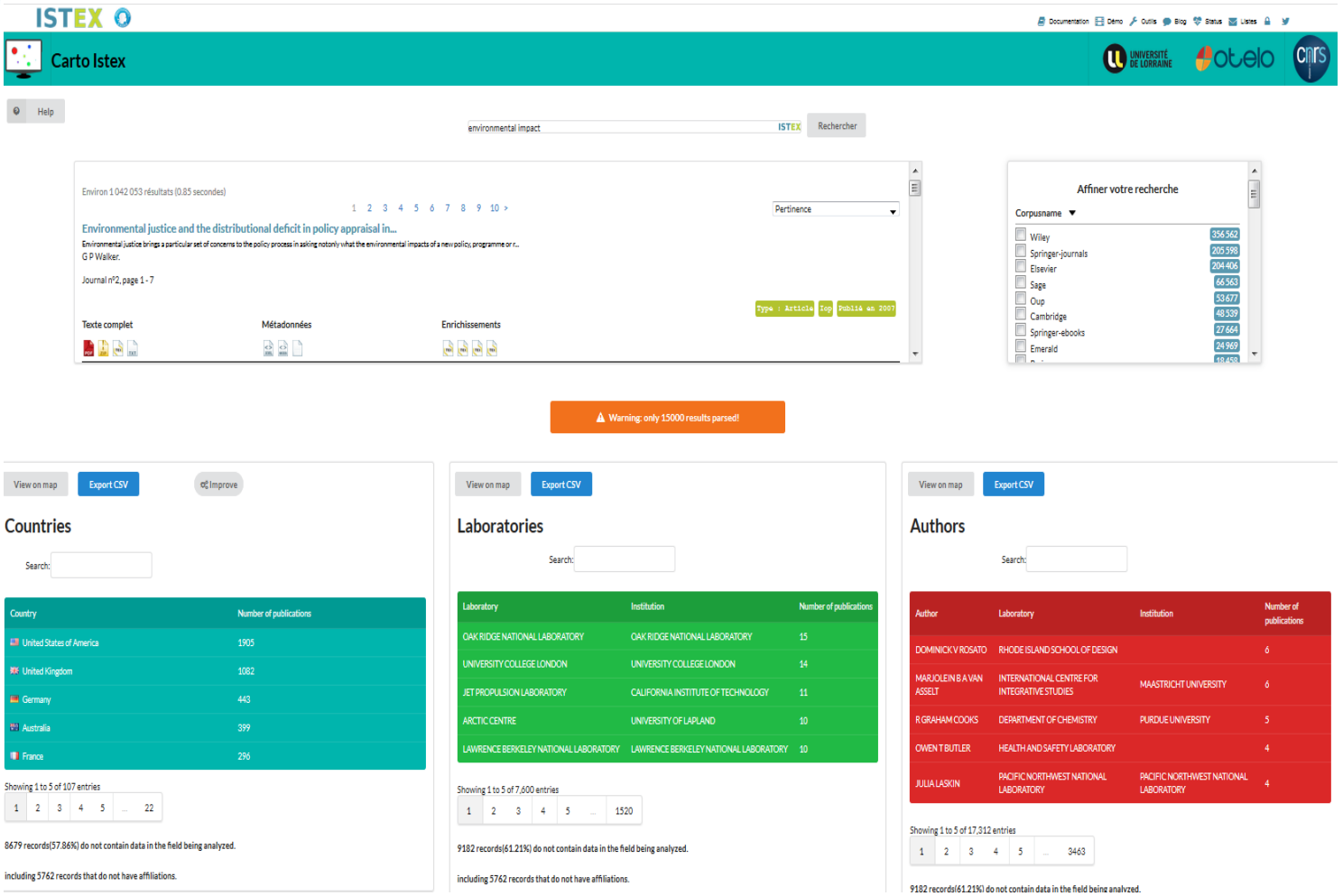
- Abstract
- ID pérenne Ark
- Score
- Nombre de mots est compté
- Full text (pdf ou txt, ou TEI)
- Métadonnées (natif XML, Mod's (format pivot), JSON)
- Enrichissements + tard ajoutés

Visa TM

Démo/CARTO ISTEX

ISTEX
L'excellence documentaire pour tous

ANR-10-IDEX-0004-02



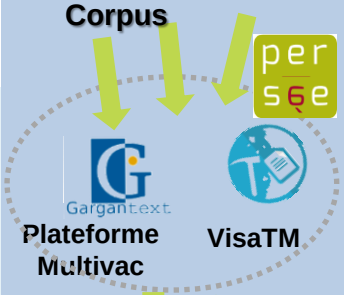
Visa TM

ISTEX
L'excellence documentaire pour tous

Mutualisation des compétences
Chaine d'ingestion des ressources
Format pivot
OCR

Catégorisation sémantique
Entités nommées

Corpus



Extraction bibliographique
Cartographie des résultats

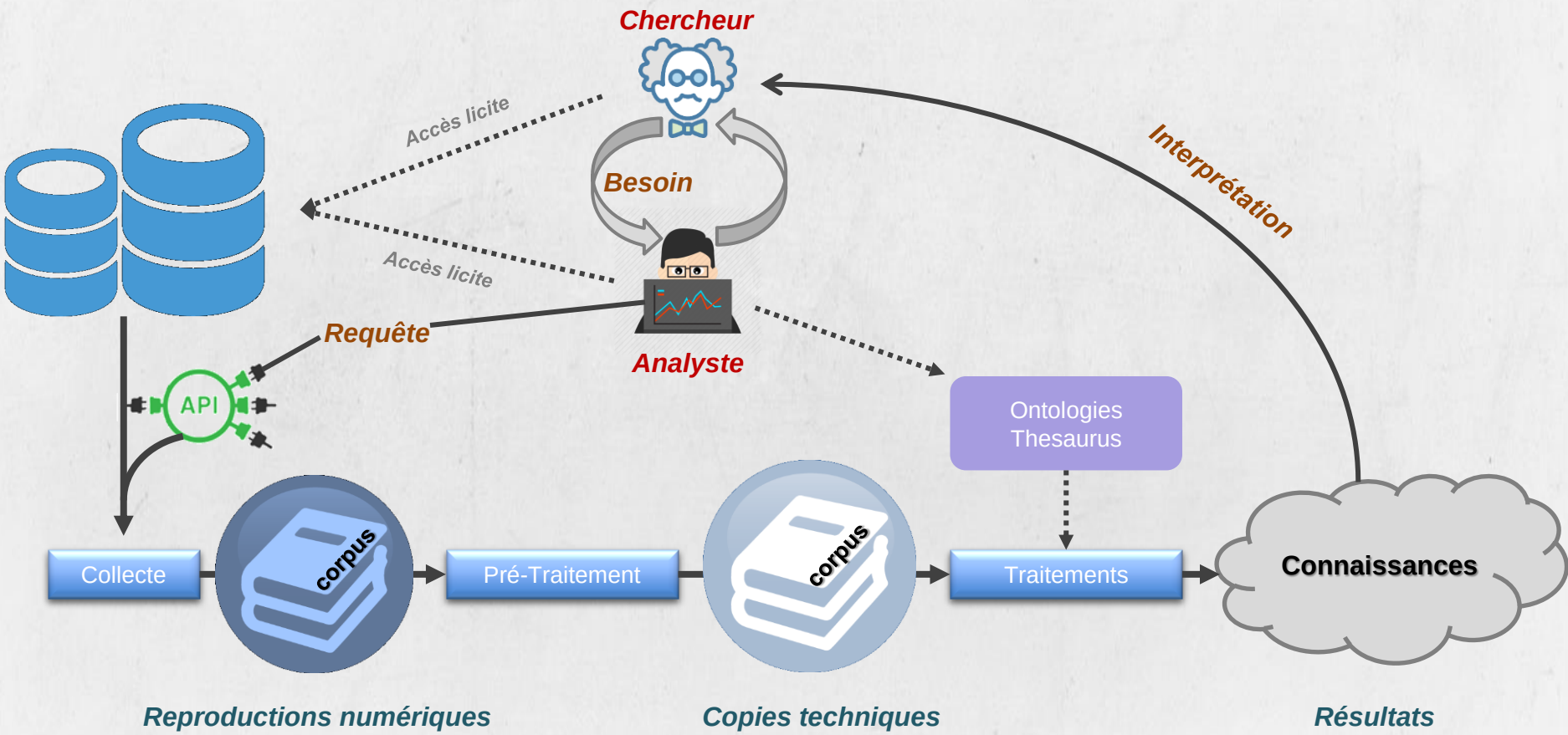
Connexion aux ENT de l'ESR & identification Renater

Recherche pluridisciplinaire TDM

- Environnement
- Médical
- Histoire
- Géologie
- SHS
- Mathématique
- Physique
- Biologie
- Informatique
- ...

A. Nazarenko et C. Nedellec

TDM Process



Visa TM

ISTEX
L'excellence documentaire pour tous

ANR-10-IDEX-0004-02

ISTEX TOUR

Les équipes ISTEX et leurs partenaires viennent à votre rencontre pour vous présenter la plateforme ISTEX, ses fonctionnalités et répondre aux questions que vous vous posez !

ISTEX.Tour@cnrs.fr

DEMO ISTEX
demo.istex.fr

GUIDE ISTEX
doc.istex.fr

BLOG TECHNIQUE
blog.istex.fr

CONTACT TECHNIQUE
contact@liste.istex.fr

COMPTE TWITTER
@istex_plateform

ISTEX
L'excellence documentaire pour tous
ANR-10-IDEX-0004-02

Visa TM

Ambitions

ISTEX
L'excellence documentaire pour tous
ANR-10-IDEX-0004-02

Etudier les conditions de production de services TDM (Text & Data Mining) à haute valeur ajoutée basés sur l'analyse sémantique à destination des chercheurs.

Doter la France d'une e-infrastructure et d'une offre de services en fouille de textes à destination des chercheurs,

en développant les synergies :

→ Entre les acteurs « Recherche » en TDM et les activités d'ingénierie et de service (pratique communautaire)

→ Avec les plateformes existantes pour enrichir leurs services (Logique contributive)

→ Avec le projet européen « OpenMinTeD » pour un effet de levier

Contribuer au développement **d'une Science ouverte** reposant sur des évolutions à la fois

- légales (Loi pour une République numérique),

- organisationnelles (Comité pour la Science Ouverte)

- scientifiques (progrès des méthodes d'analyse textuelle, Linked Open Data, analyse sémantique basée sur des terminologies et ontologies)

Visa TM

Organisation

ANR-10-IDEX-0004-02

Des partenaires

- Un établissement de recherche
- Un opérateur IST
- Une Université (Montpellier)

Un financement

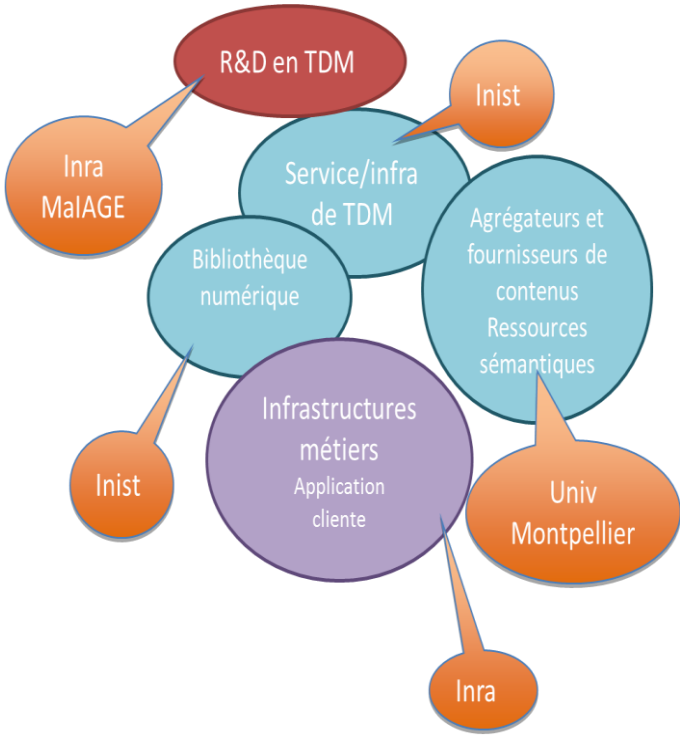
Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation

Un comité de pilotage

…sous l'égide du CoSO dans le cadre de sa stratégie « Open Science »

Un chef de projet

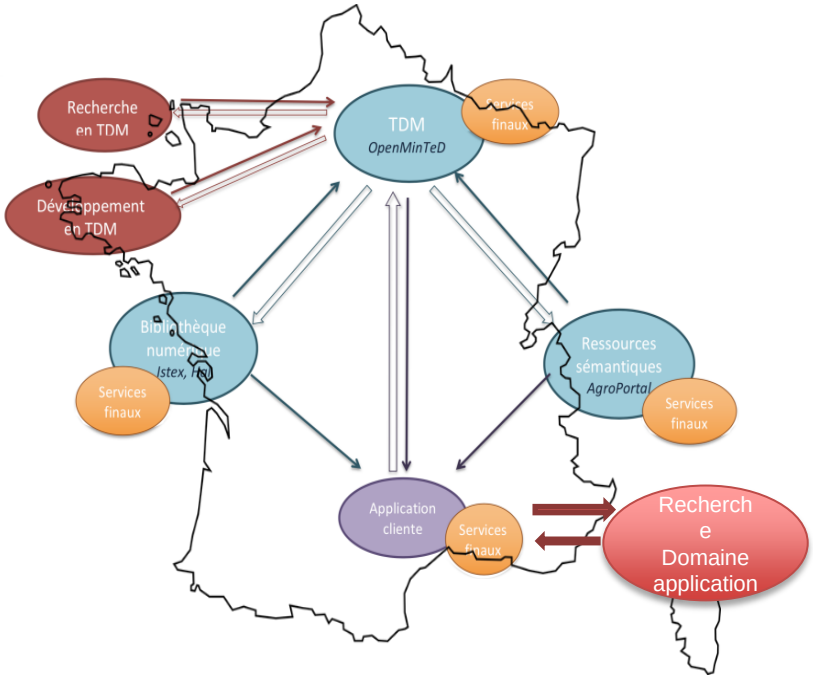
Claire Nédellec
INRA – Unité MaIAGE



La démarche

ISTEX
L'excellence documentaire pour tous
ANR-10-IDEX-0004-02

Un pas vers une généralisation des approches TDM dans les activités de recherche



1. **Analyser** les opportunités et les besoins des différents acteurs
2. **Tester** la faisabilité technique des interconnexions entre plateforme TDM, bibliothèques numériques et portails de ressources sémantiques = 3 piliers d'une infra TDM
3. **Démontrer** l'utilité au travers de 3 applications pilote combinant TDM, corpus documentaires et ressources sémantiques.
4. **Proposer** une infrastructure technique et humaine, ouverte et pérenne proposant une offre de services en fouille de textes et de données dans le contexte français

Trois chantiers

ANR-10-IDEX-0004-02

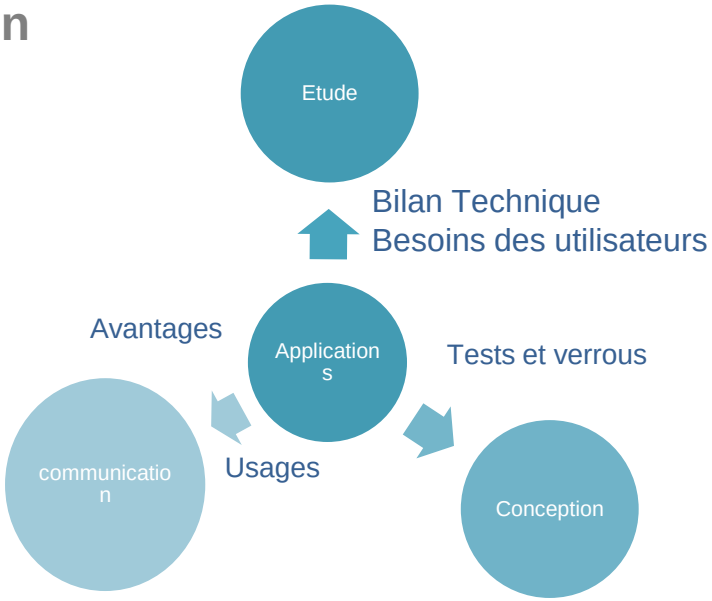
Le projet VISA ™, une opportunité française dans un cadre européen

Volets

Étude (*Inist*) : Analyse de l'existant et besoins, articulations des stratégies nationales et des organismes, scénarios techniques et organisationnels, feuille de route

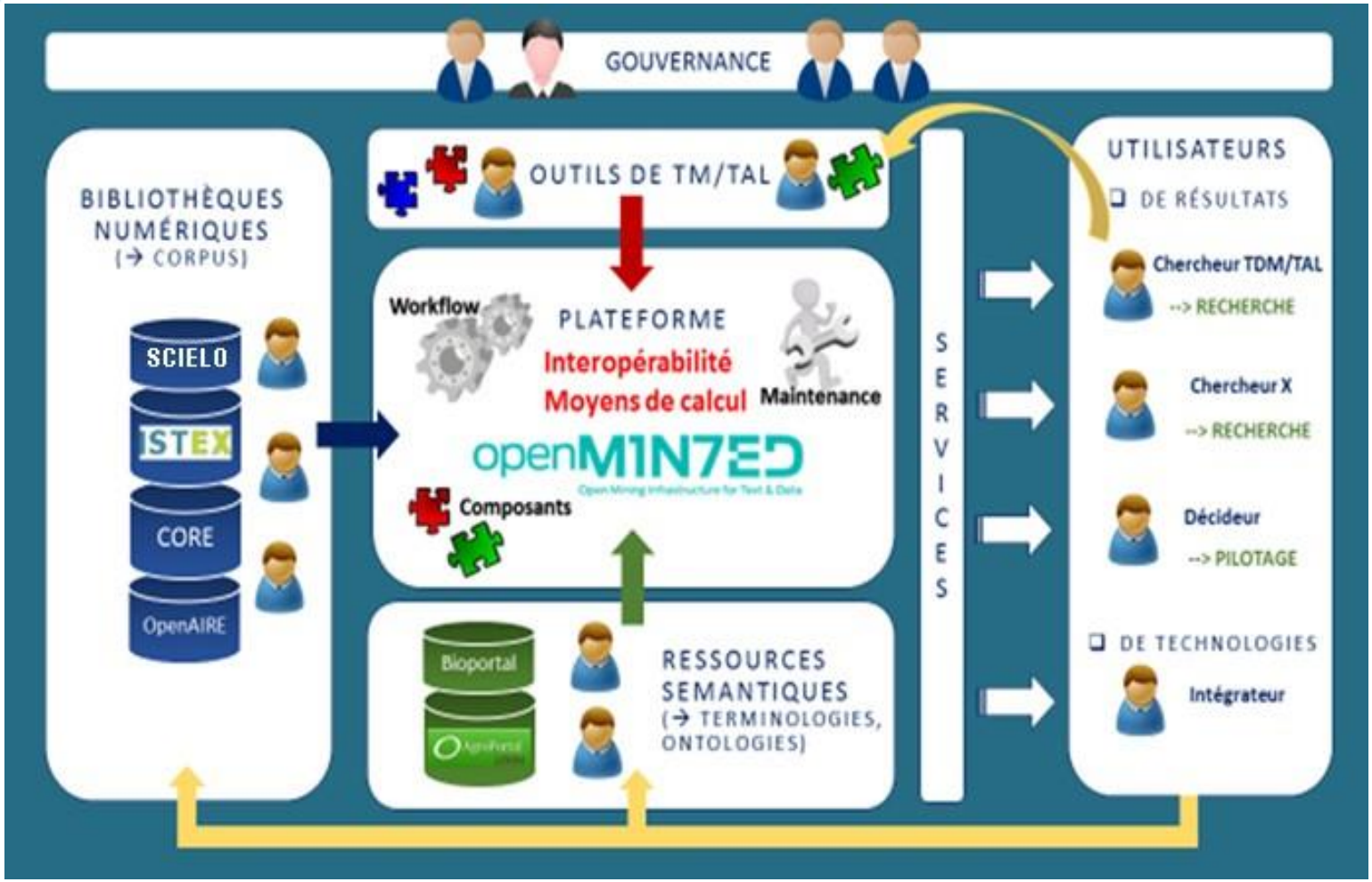
Conception (*Inra*) : Utilisabilité et faisabilité du couplage des 3 piliers (OMTD, ISTEX, Agroportal)

Démonstrateurs (*Inra*): illustrer la facilité de déploiement et la qualité des résultats répondant à des besoins identifiés : IST, Sciences de la Vie, éditeur de *workflow*



E-infrastructure cible

ISTEX
L'excellence documentaire pour tous
ANR-10-IDEX-0004-02



Visa TM

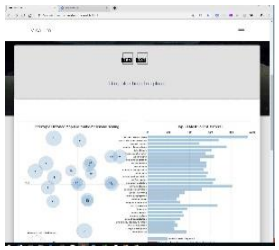
Connexion OpenMinted à ISTEX
- l'exploration de corpus -

ANR-10-IDEX-0004-02

Outil d'aide à la construction et l'exploration d'un corpus de documents scientifiques issu de ISTEX, à l'intention de toute personne qui désire caractériser et affiner un corpus en s'appuyant sur une représentation thématique de l'information, (cartographies de thèmes), ainsi que sur des statistiques descriptives fondées sur les métadonnées bibliographiques.

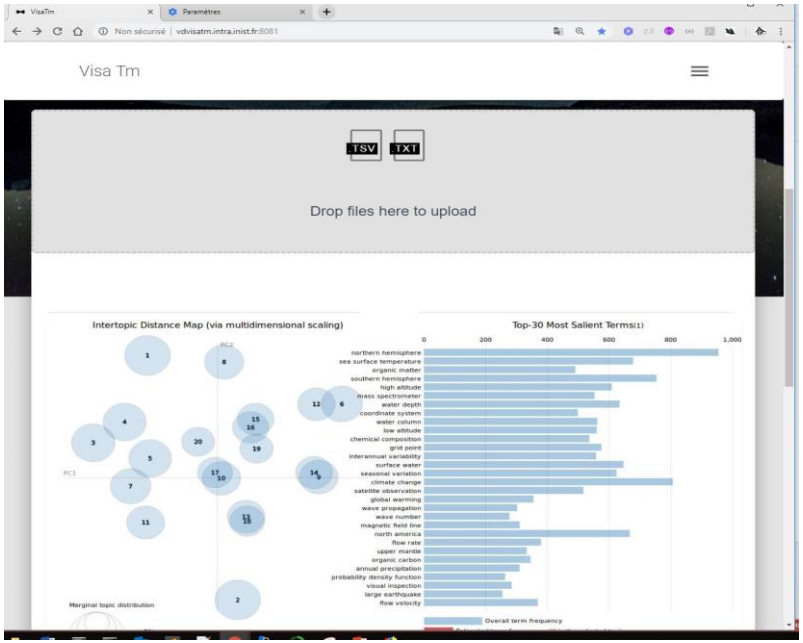
Analyse corpus

Extraction corpus



Visualisation information

Analyse thématique **5000 documents** de Geosciences



Visa TM

Une étude

ANR-10-IDEX-0004-02

1 Pour décrire le contexte

Quels acteurs ? : producteurs (de contenus, d'outils), utilisateurs (chercheurs TAL ou non, décideurs, intégrateurs, développeurs…), communautés qui gravitent autour, « accompagnants » (cf. rôle des documentalistes) etc.

pour déterminer des **compétences, des missions et des profils** nécessaires et des profils

Quels outils ?

Une base: le projet OpenMinTED

Des outils à intégrer : éventail de l'existant et analyse des fonctionnalités, de l'interopérabilité, des coûts d'utilisation (libre/payant), …

Quels besoins?

=> Questionnaire en cours de finalisation

2 Pour proposer un scénario d'organisation : infrastructure et RH

3 Pour proposer une feuille de route

Visa TM

ISTEX
L'excellence documentaire pour tous

ANR-10-IDEX-0004-02

openMINTED
Open Mining Infrastructure for Text & Data

Visa TM

open MIN7ED , le projet

ISTEX
L'excellence documentaire pour tous
ANR-10-IDEX-0004-02

Open Mining INFrastructure for Text and Data

Projet H2020 d'infrastructure européenne de text-mining Open Source

Coordination par l'Université d'Athène

16 partenaires – 1er juin 2015 à fin mai 2018

→ Reposant sur une analyse de l'existant et du besoin

→ Connection aux infrastructures existantes dont bibliothèques numériques

→ Catalogue riche de composants TDM interopérables, pour la composition, la réutilisation, et l'exécution de workflows de traitements TDM

→ Orienté services

→ Mettre le TM à la portée de tous (interface user friendly), sans être un specialiste du domaine

Opportunité Capitaliser sur ce projet

Bénéficier de l'engagement des communautés thématiques

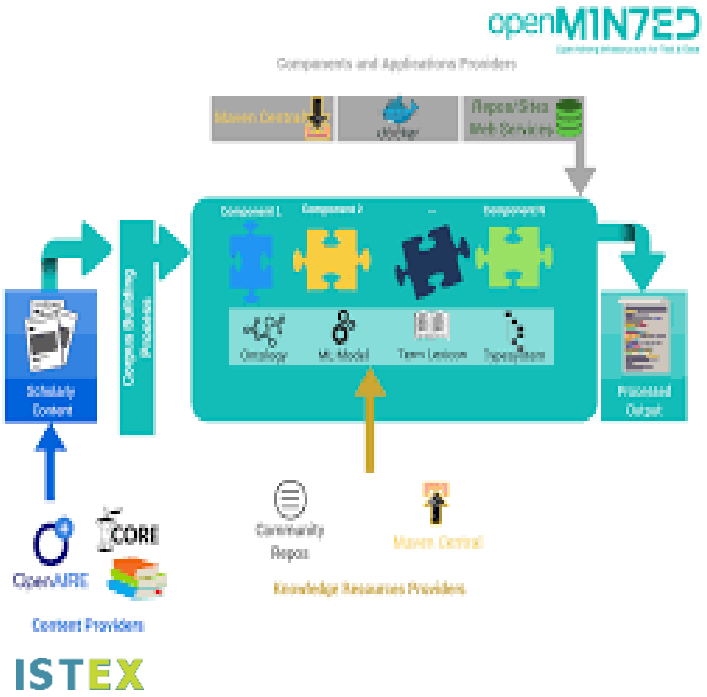
Visa TM

openMINTED, pour faire quoi ?

ISTEX
L'excellence documentaire pour tous

ANR-10-IDEX-0004-02

Une Infrastructure ouverte de *text-mining* où les chercheurs peuvent découvrir, créer, partager et réutiliser des logiciels, des documents et des ressources pour le TM, TAL, EI, etc.à partir de source scientifiques (publications)

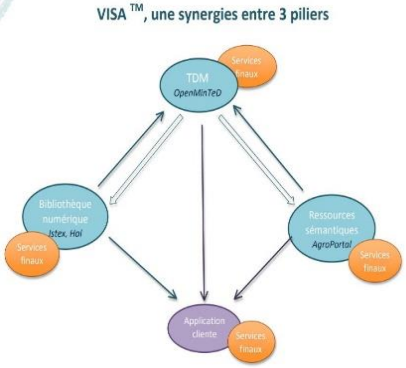


- **Des corpus** — Collection de documents
- **Des applications** — Traitement text mining avec lequel un corpus peut être traité sans configuration
- **Des composants** — Outils linguistiques unitaires. Les applications sont construites par assemblage et configuration de composants
- **Des ressources** — Connaissances formalisées utilisées par les outils : lexique, ontologies..

Visa TM

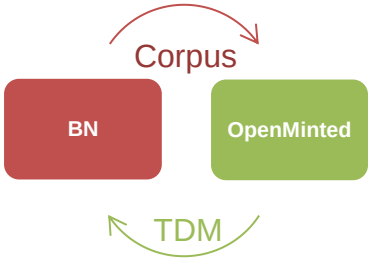
et les bibliothèques numériques

ANR-10-IDEX-0004-02



Les bibliothèques numériques comme un des piliers de la e-infrastructure

Apport mutuel



- Construire des corpus au plus près des outils de TDM
- Interrogation multi source CORE ISTEX OpenAIRE
- Possibilité de télécharger des corpus existants (Format documentaire natif : XMI, TXT, PDF ; sinon convertisseur TEI …)
- Capacité d'identification (OMTD-SHARE Metadata corpus) et de partage de corpus
- Possibilité d'extension de bibliothèques au grès des partenariats

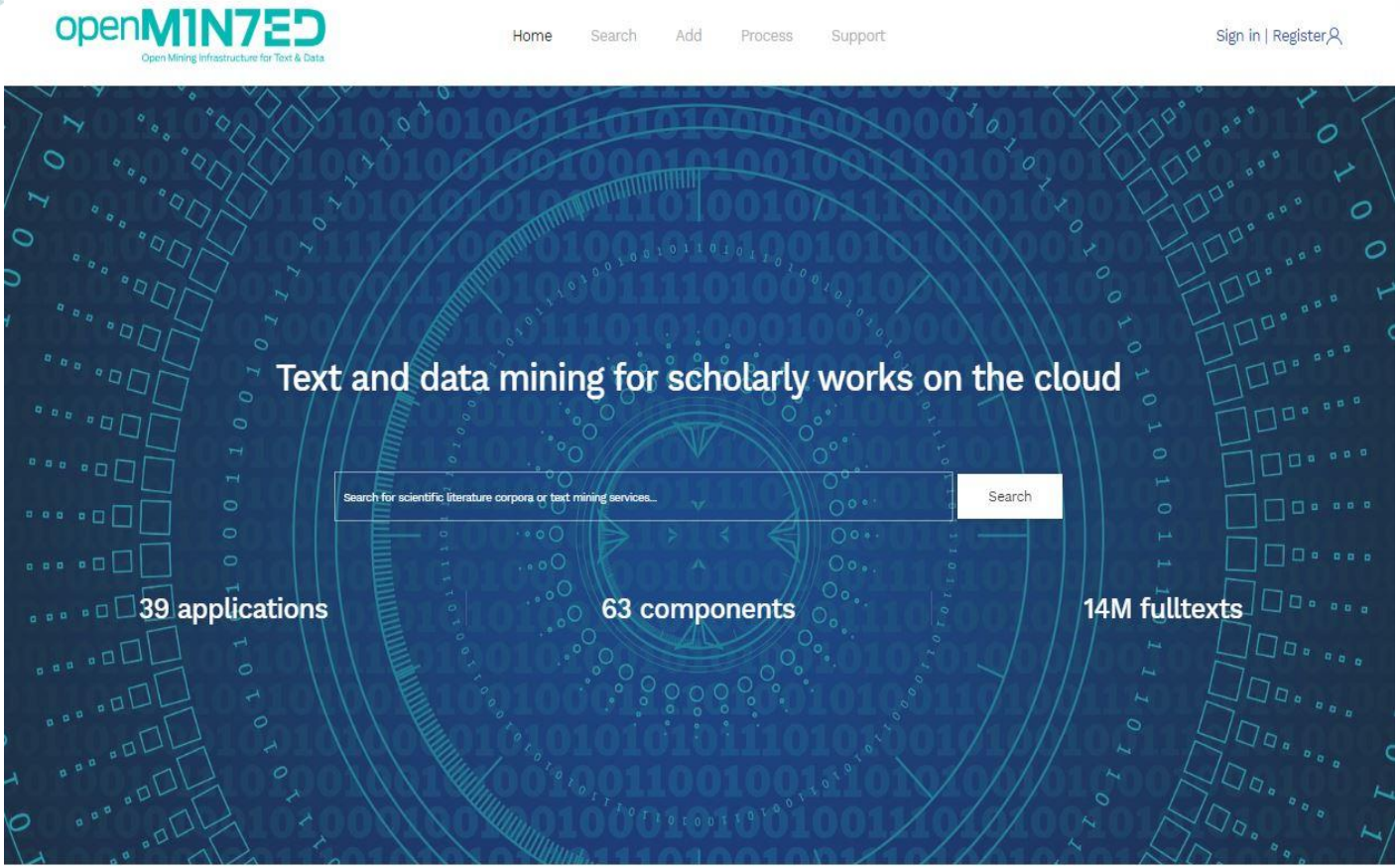
Un corpus : Une collection structurée de données (textuelles, audio, vidéo, multimodales / multimédias, etc.) généralement de taille importante et sélectionnées selon des critères externes à ces données (taille, type de langue, type de producteur ou public visé, etc. .) qui représente de manière aussi complète que possible un objet de l'étude.

Visa TM

https://services.openminted.eu/home

ISTEX
L'excellence documentaire pour tous

ANR-10-IDEX-0004-02



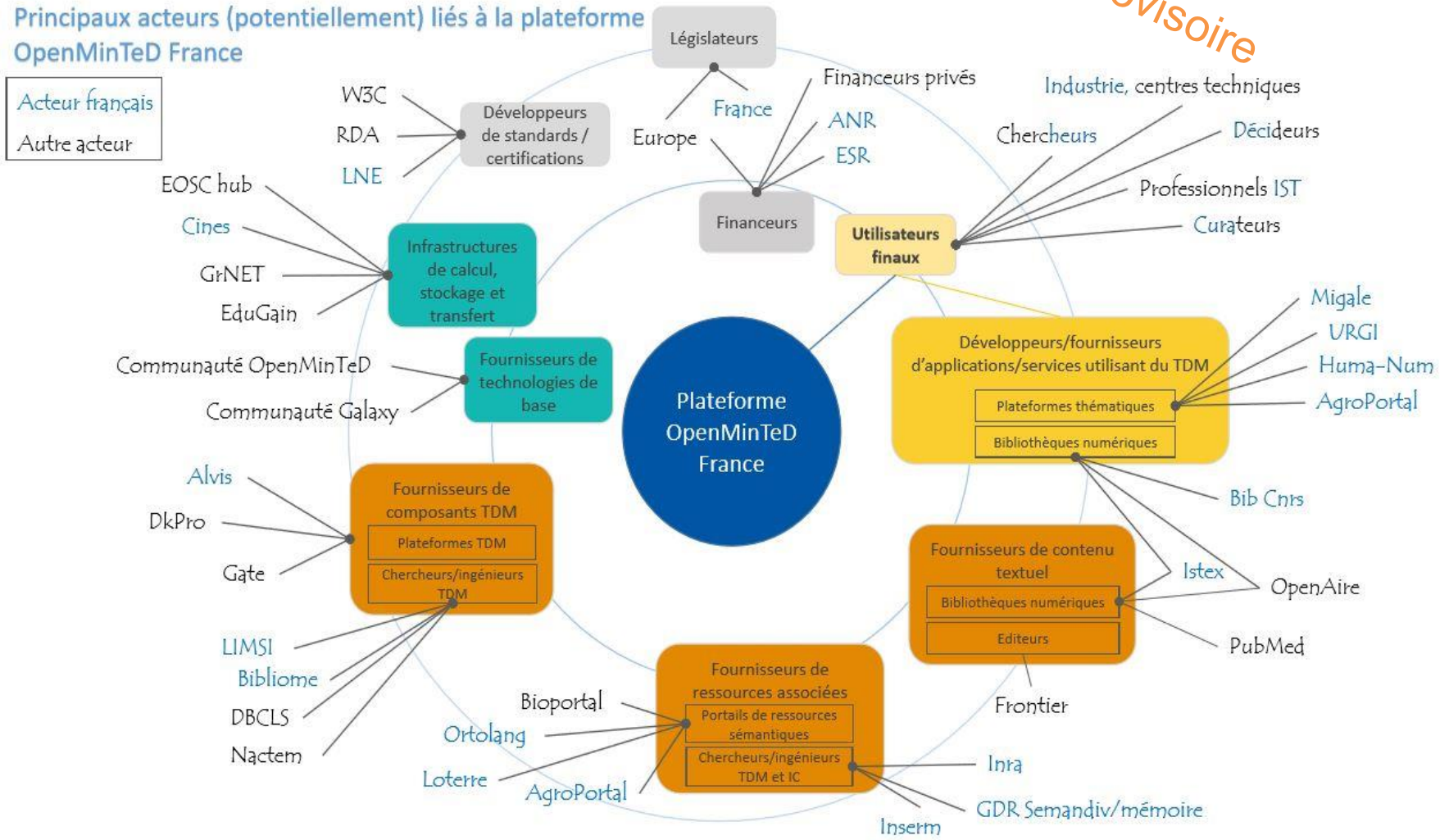
ISTEX
L'excellence documentaire pour tous
ANR-10-IDEX-0004-02

Le rôle du documentaliste

Les acteurs

ANR-10-IDEX-0004-02

Version provisoire



Quel rôle pour le documentaliste?

ANR-10-IDEX-0004-02

La science ouverte bouscule les pratiques et les métiers de la documentation

Aujourd'hui, l'arrivée du **TDM** → nouveaux défis

- Data Mining en forte relation avec les données de recherche (des sensibilisations, des formations, une nécessité d'accompagnement des chercheurs)
- Text Mining en forte relation avec les bibliothèque numériques et les ressources sémantiques (terminologies, ontologies, web sémantique)

Poursuivre le soutien et l'accompagnement du chercheur: outils, ressources, aspects juridiques, …

- Des formations à prévoir sur de nouvelles compétences et des profils de poste en évolution
- La nécessité d'un travail collaboratif en équipe : chercheur, documentaliste, informaticien

- Développement de services de soutien
- Elaboration d'étude de cas, de guide de bonne pratique
- Encadrement et suivie des expériences menées par les chercheurs

08/10/2018 Fredoc2018 - ISTEX et OpenMintedD, vers une infrastructure de text-mining 32