

**OUVRIR
LA SCIENCE !**

Journées Nationales de la Science Ouverte

Fouille de texte, un écosystème en évolution



Claire Nédellec



MESRI, Paris
6 décembre 2018

Fouille de texte, un enjeu scientifique majeur

Plus de 60 millions d'articles, 2.5 par an
160 millions de documents scientifiques
indexés

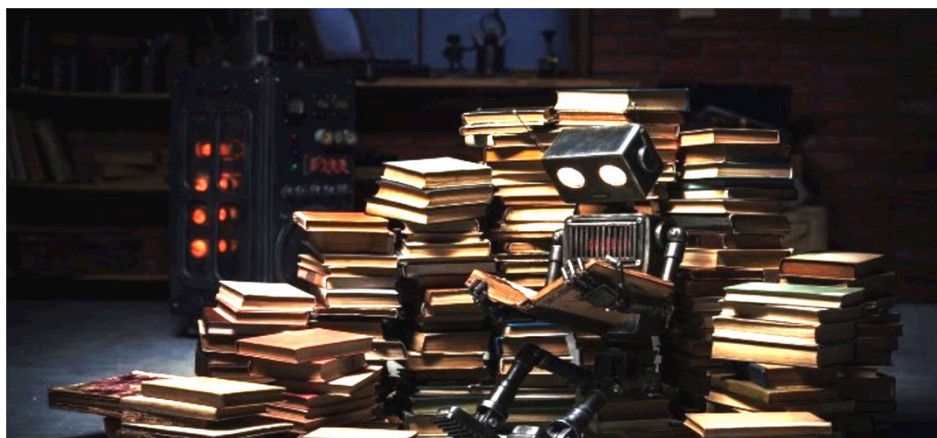
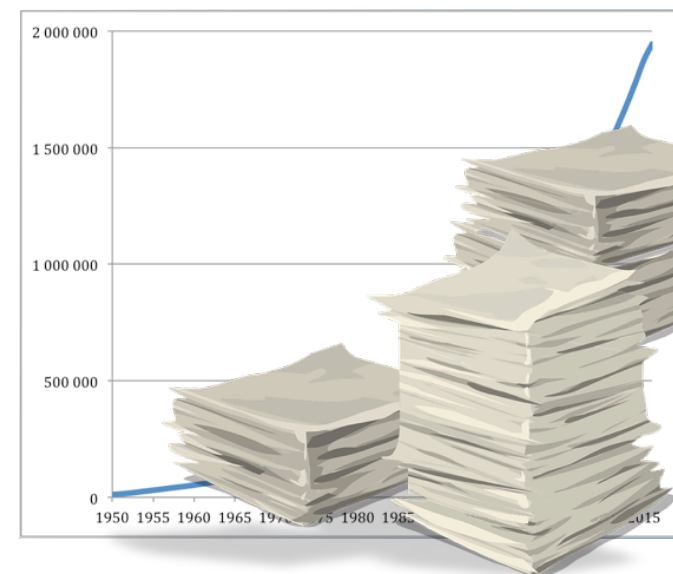
*Orduña-Malea et al.,
Scientometrics 2014*

50% des articles ne sont jamais lus
90% des articles ne sont pas cités
80% des articles cités ne sont pas lus

*The STM
report, 2015*

*Lokman I. Meho, the
rise and rise of citation
analysis, 2007.*

*Simkin & Roychowdhury.
Read before you cite!,
2002*



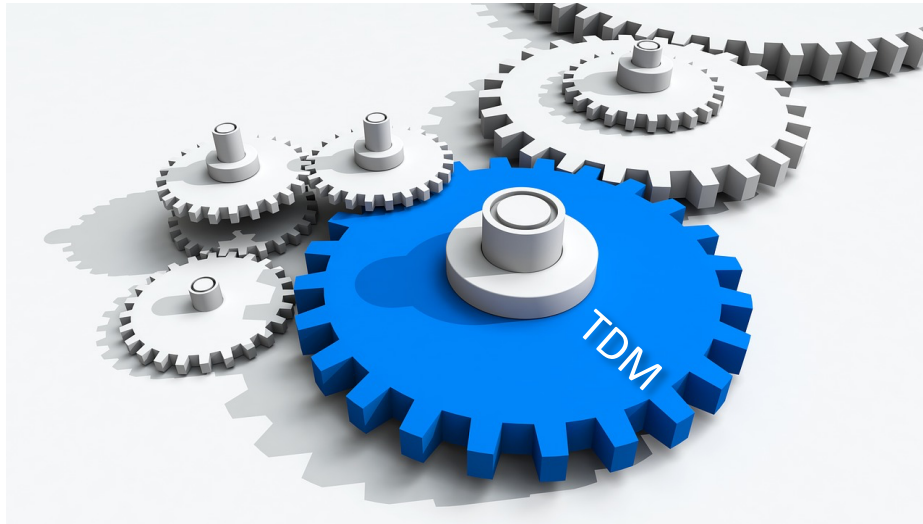
Exploiter les données de la recherche

Donner du sens aux données textuelles
Transformer une donnée non structurée
en donnée structurée, manipulable par
un ordinateur

Intégrer le TDM scientifique
au cœur de l'activité du chercheur
non spécialiste



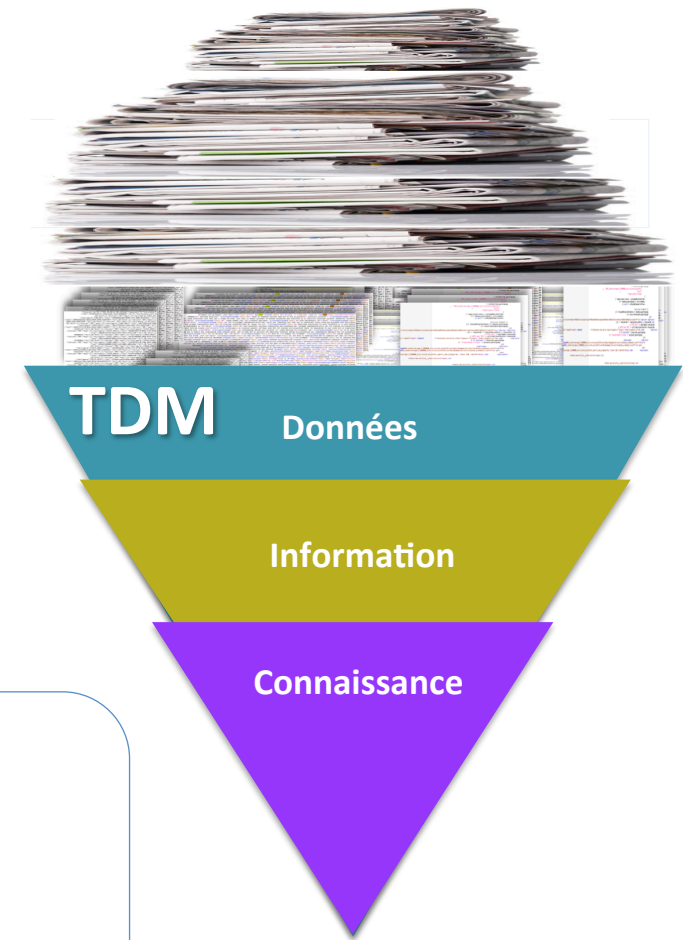
Le TDM intégré dans le traitement des données et des informations



Des traitements spécifiques pour des données pas comme les autres

De nouvelles voies pour la découverte de connaissances
Hargreaves Report (Digital Opportunity, 2011)

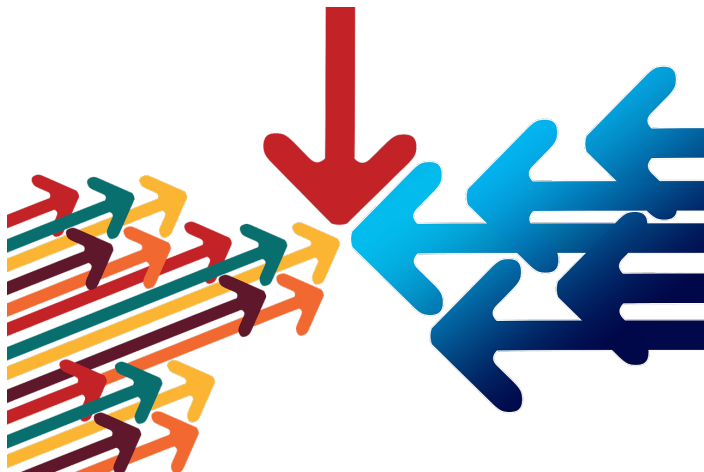
- Pour un bénéfice immédiat pour la recherche
- Un accélérateur d'innovation
- Répondre à l'évolution de la recherche vers plus de transversalité.



Le TDM pour la science, des enjeux de recherche et d'ingénierie

Convergence

- Maturité des technologies de TDM et du web sémantique
- Efficacité et disponibilité des moyens de calcul
- Accessibilité des bases bibliographiques
- Standardisation des accès et des représentations, sécurité juridique grandissante
- Développement des infrastructures de recherche



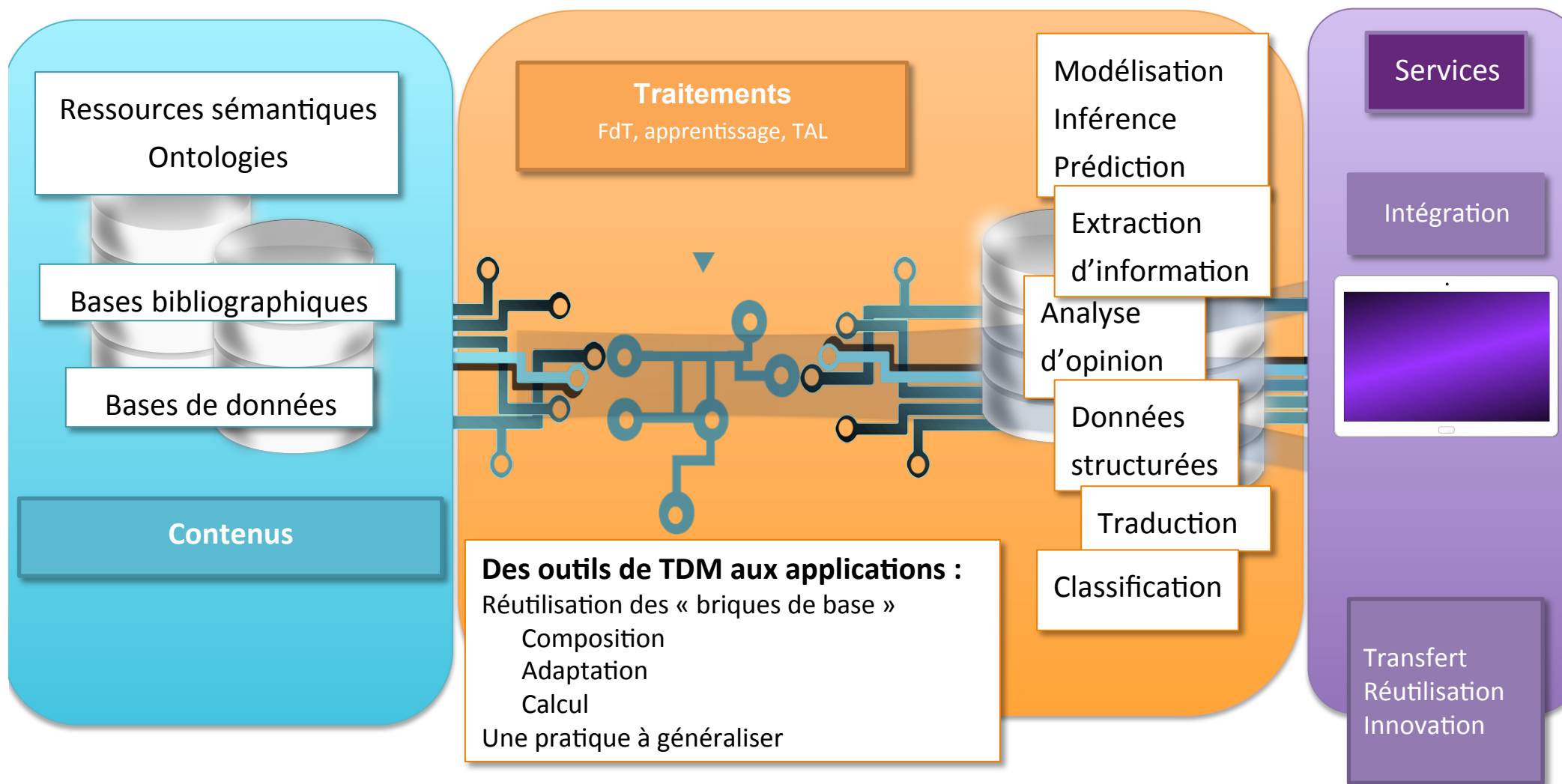
Un paysage fragmenté

fournisseurs de contenus, chercheurs en TDM,
infrastructures de calcul, services
Des outils très nombreux et hétéroclites
pour traiter la diversité

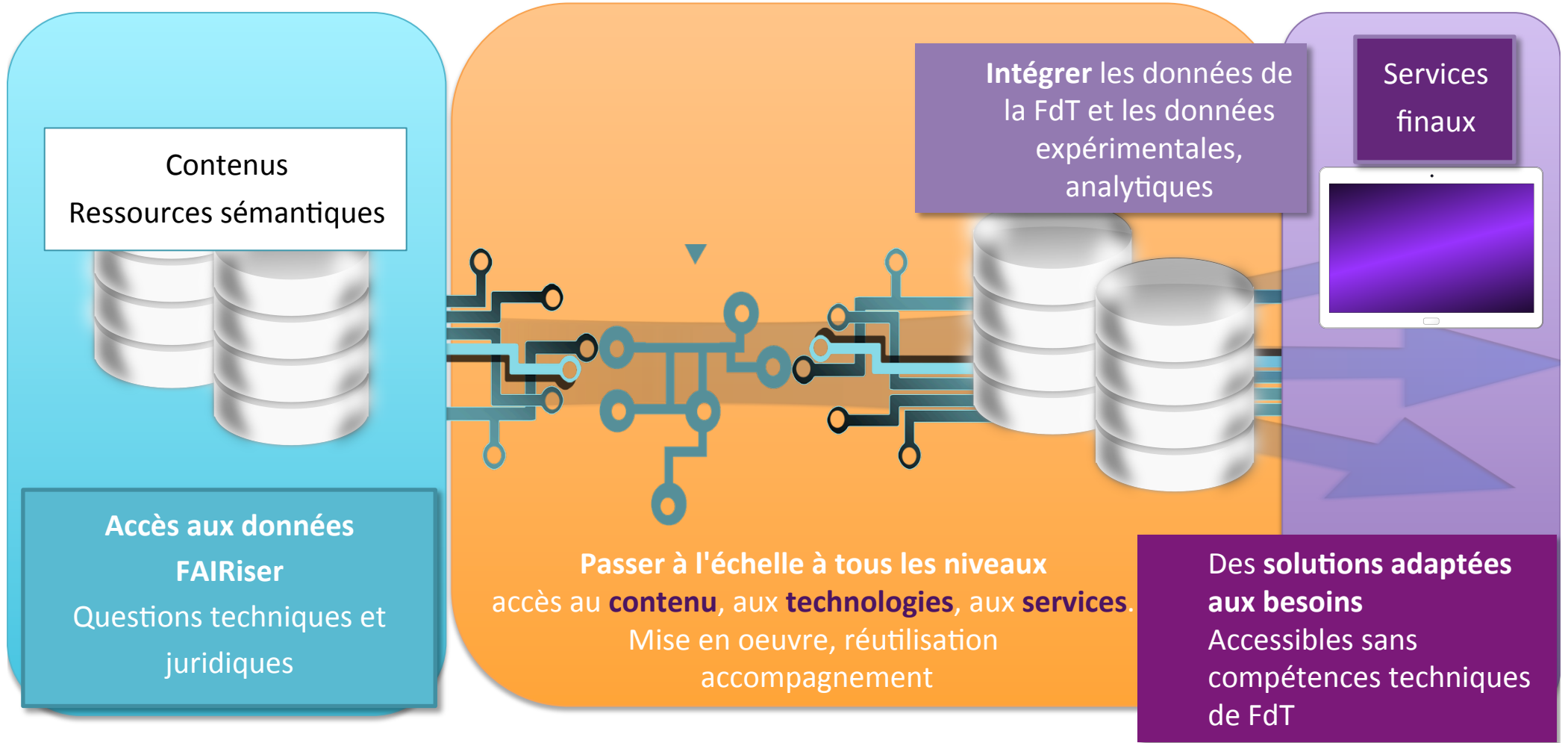
De nouvelles questions économiques, juridiques,
techniques et organisationnelles

Pour des services à la recherche





Fouille de texte pour *la recherche*



Stratégie partagée : rationaliser – mutualiser

S'inscrire dans une stratégie nationale et européenne de développement d'infrastructures de recherche mutualisées et coordonnées, de données et de traitements

(1) Les Données

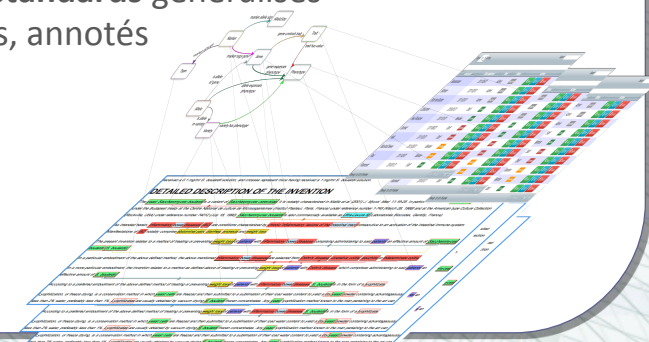
Pour

- **Rationaliser et mutualiser** l'accès aux **sources documentaires et sémantiques**.
- Réduire le coût d'ingénierie de la conception de corpus
- Rendre l'accès ouvert, transparent, adapté aux besoins et fiable (qualité, sécurité juridique, caractérisation)
- Intégrer les données, résultats de FdT avec les autres données

Comment

- Changer les pratiques de publication
- Rendre les publications accessibles et réutilisables dans des **formats standards** généralisés
- **Agréger**, centraliser, partager et réutiliser des corpus bruts, prétraités, annotés
- Développer et partager des **ressources et modèles sémantiques** spécialisés
- **Combiner et réconcilier** les données par des ontologies de référence

...



Stratégie partagée : rationaliser – mutualiser

S'inscrire dans une stratégie nationale et européenne de développement d'infrastructures de recherche mutualisées et coordonnées, de données et de traitements

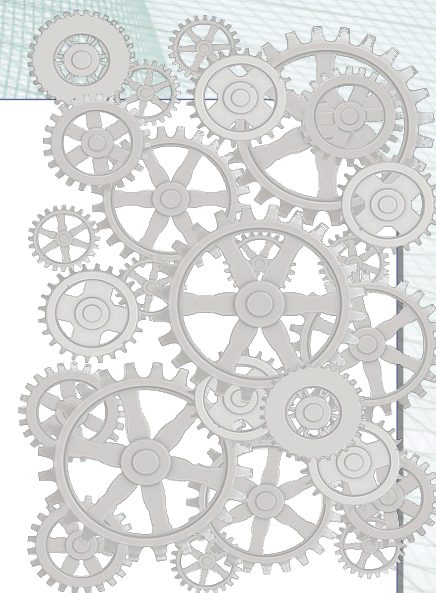
(2) Les Traitements

Pour

- **Réutiliser, combiner** les outils de TDM pour les adapter aux besoins (aujourd'hui, des milliers d'outils et des plateformes techniques).
- Créer des environnement favorables à **l'expérimentation et la reproduction**

Comment

- **Plateformes** de traitement et service
- **Interconnecter** durablement les sources de données, les traitements et les services
- **Calcul** haut débit
- **Interopérabilité** des composants de TDM
- Fournir aux utilisateurs sur leurs postes de travail des **modes d'interaction adaptés** à leurs profils
- Pour plus de finesse et d'**adaptation au besoin**, mieux intégrer l'apprentissage automatique et les ressources sémantiques. Gérer le compromis généricité-adaptation / coût-valeur ajoutée.



Concevoir une solution académique, par les académiques, pour les académiques ?

Inscrire le TDM
dans la **stratégie nationale, européenne et internationale
des données et services**

Solutions industrielles et académiques : complémentaires et supplémentaires

- Mutualisation et partage des résultats de la recherche en TDM pour la recherche académique : une **problématique collective**
- Des solutions de TDM académiques pour **des besoins non rentables** pour les entreprises du secteur : dans les domaines de production de la connaissance, non marchands
 - Coût de l'ingénierie documentaire, expertise du domaine, maintenance du service
 - Relevant des missions des organismes de recherche

Et parallèlement sont un levier d'innovation

Ainsi, l'accès aux publications

Le transfert des prototypes vers des produits

Modèles mixtes à inventer

Services des PME innovantes exploitant des ressources libres



Journées Nationales de la
Science Ouverte

Fouille de texte, un
écosystème en évolution