# Retour sur l’atelier TDM – CARIST 2017

* **Jean-Marie Pierrel (Université de lorraine)** fait un rappel introductif sur le TDM. Il situe son avènement sur une échelle de temps démarrant vers 1994 avec le concept de data mining suivi en 2000 par celui de text mining. Il orientera explicitement son exposé essentiellement du côté du text mining en précisant les différents étapes qui lui sont nécessaires : accès et collecte des données, ingénierie documentaire, ingénierie du langage, fouille et visualisation des résultats. L’analyse fine des données étant du ressort du chercheur.

Dans l’enchaînement de ce processus, il apparaît aujourd’hui assez clairement qu’existe un besoin d’**accompagnement de l’usager**, passant par une étape de **formation, à la fois du chercheur et du professionnel de l’IST**.

ISTEX, autant en tant que réservoir de constitution de corpus qu’en tant que « bac à sable » pour des expérimentations et cas d’usage, est un outil fort intéressant aujourd’hui. La V2 devrait être tournée en particulier vers la fouille de textes.

* **Présentation du chantier d’usage ISTEX par Patrice LOPEZ de l’équipe ALMAnach** (anciennement Alpage) de l’INRIA (porteur = L. Romary) Sont abordés plusieurs travaux et outils : Grobid (GeneRation Of BIbliographic Data) qui permet de convertir le pdf en TEI, Keyterm qui permet la désambiguïsation avec Wikipédia/Wikidata, (N)ERD (Names-Entity Recognition and disambiguation).

L’extraction d’entités nomenclaturées sur des corpus de HAL a permis par exemple un travail d’extraction d’entités astronomiques et leur mise en lien avec une base de données d’objets astronomiques.

Un travail du même ordre a permis un lien entre AUréHal et Idref (ABES).

Sur ISTEX l’industrialisation d’annotateurs trouve une application dans l’extraction d’entités représentatives de quantités et leur normalisation par rapport à des unités internationales.

* **Présentation de Terre-ISTEX par Eric Kergosien** (Laboratoire Gériico-Université Lille 3) : Identification et analyse des terrains d’étude dans les corpus ISTEX. La thématique traitée est le changement climatique. Il s’agit de déterminer les fronts de recherche en fonction des territoires d’étude, des frontières disciplinaires et en tenant compte aussi de l’extraction d’entités temporelles (à l’aide de l’outil HeidelTime/GitHub). Une application possible : la valorisation des données de la recherche - Vador (31 mai 2017, Toulouse), 4Humanities…
* **Présentation de Jacques DUCloy sur lorexplor : « L’excellence documentaire pour tous ? »**

L’application permet une exploration rapide de corpus en associant à la fouille un réseau wicri, des données Wikipédia et pourrait être assimilée à un Wikipédia de la recherche. Aujourd’hui on répertorie une moyenne de 10000 documents par sujet exploré, on peut en espérer jusqu’à 20000 ultérieurement.

Un intérêt non négligeable : la formation, la collaboration…

* **Une vision prospective des textes scientifiques par Guillaume Cabanac** de l’IRIT de Toulouse.

Ici on interroge le texte scientifique au sens large (pas seulement les revues scientifiques) et on met en œuvre l’interdisciplinarité par des liens avec la sociologie, la psychologie dans le cadre d’études scientométriques. Exemple : utilisation du TDM pour la détection d’éponymes pour aller au-delà de la simple citation scientifique, opérations « SocioNoël » ou encore « l’ESR en peinture »…

*Des projets de recherche en TDM intéressants donc, mais encore un manque de visibilité claire de leur application au quotidien pour le chercheur lambda dont on peut se demander comment il peut s’approprier ces outils pour les utiliser de façon optimale…*

*Une rupture certes mais pas tant au niveau des techniques appliquées qu’au niveau des changements au niveau des métiers qu’une formation adéquate devra accompagner. Le professionnel de l’IST devrait y trouver une place mais à condition d’adapter ses acquis à de nouvelles exigences et de nouveaux objets.*