Note EPRIST sur le text et data mining :

Le TDM comme outil innovant de recherche scientifique

Dans le cadre de la préparation du projet de loi sur le numérique, EPRIST a rassemblé des exemples de pratiques au sein des équipes de recherche des organismes français de recherche, afin de montrer l'importance du TDM comme outil de traitement de données à des fins de recherche.

Les échanges réalisés avec les scientifiques concernés à cette occasion ont fait ressortir les opportunités des usages de ces outils, tout comme les difficultés rencontrées par les chercheurs (freins juridiques à l'usage du TDM en l'absence d'une exception pour les travaux de recherche). Ces freins pénalisent actuellement la France dans la compétition internationale, à l'heure où la recherche est de plus en plus évaluée sur sa capacité à innover et, plus généralement, sur ses impacts sociétaux.

Les pratiques actuelles :

Les équipes de recherche développent partout de nombreuses applications de TDM qui exploitent les textes intégraux d'articles scientifiques dans leurs champs de recherche et dans les domaines connexes. Certains projets scientifiques basent également leurs travaux sur des traitements croisés de données bibliographiques, de brevets, de magazines professionnels (recherche finalisée), de la presse généraliste (liens sciences-société).

De leur côté, les professionnels de l'information scientifique et techniques (IST) utilisent le TDM dans le cadre des services d'analyse de l'information à haute valeur ajoutée, réalisés pour les scientifiques et les décideurs en appui à la stratégie scientifique de leur organisme, notamment en amont des processus d'innovation et de valorisation économique.

De fait, depuis toujours, l'analyse de corpus de publications scientifiques, de références bibliographiques, ou de jeux de données de tous types, est au cœur du métier et des pratiques des chercheurs. En premier lieu, ces analyses permettent d'éviter de réinventer ce qui existe, et de bénéficier des apports de ses pairs. Ensuite, il faut extraire ce qui est utile dans ces corpus d'informations, composés de sources multiples, pour mieux croiser, réinterpréter et réintégrer ces informations dans d'autres contextes. Dans le processus de recherche, le TDM est un outil « moderne » essentiel, adapté à la masse croissante de données polymorphes (BigData), pour exploiter les données et produire de nouvelles connaissances. Le TDM est un outil de découverte incontournable au XXIème siècle, à l'heure du Big Data, du Libre Accès, dans toutes les disciplines.

Les scientifiques travaillent aujourd'hui à relever les grands défis sociétaux que sont la connaissance et préservation de l'environnement, l'adaptation au changement climatique, la santé humaine, l'alimentation, l'énergie… Pour traiter ces sujets complexes et souvent interdépendants, ils conduisent des travaux basés sur des approches à la fois disciplinaires et transdisciplinaires, et ceci dans un contexte de compétition mondiale féroce. Limiter l'usage du TDM par les scientifiques français présente un handicap majeur à la fois en termes de périmètre (objets multiples), de volumétrie (très grande quantités d'informations) et de rapidité de traitement (démultiplicateur) ;

l'enjeu est d'autant plus grand que leurs concurrents à l'échelle européenne et internationale bénéficient aisément de ces outils.

Notons enfin que les techniques de TDM s'appuient sur des outils innovants (logiciels, moteurs de recherche, technologies pour la collecte et le traitement massifs de données) que la Recherche (Inria, CE, CNRS…) et les entreprises françaises contribuent à développer. C'est un des grands secteurs d'innovation potentiel actuel. Nous priver de la possibilité de concevoir, tester ou utiliser ces outils nous exclue d'un marché d'avenir et nous retarderait considérablement dans la compétition mondiale que constitue aujourd'hui l'Economie numérique.

Exemples d'usages par les scientifiques :

Les usages sont donc nombreux et touchent toutes les disciplines. Sont rassemblés ici quelques exemples pour illustrer notre propos.

Au CEA, la pratique du TDM est très courante tant en physique des particules, en physique nucléaire ou en astrophysique (cf. travaux du Centre de données astronomiques de Strasbourg : http://cdsweb.u-strasbg.fr/about). La mise en relation d'archives interopérables et d'outils intelligents pour leur analyse est à la base de projets européens, comme The European Virtual Observatory, http://www.euro-vo.org/ , ou EUROPLANET, http://www.europlanet-eu.org/ . C'est également le cas en physique de la matière condensée par exemple qui est à l'origine de toutes les innovations sur les matériaux, les nouvelles technologies de l'énergie, de la santé ou de la communication. Enfin, le principe même de pouvoir exploiter des « archives » de nature extrêmement variées est un des pilier des sciences du Climat et de l'environnement, comme le montre le travail du LSCE dans la reconstitution des climats du passé.

Dans d'autres domaines, l'usage du TDM est utilisé par des équipes du CEA pour le traitement automatique des langues, notamment pour la constitution, la visualisation et l'analyse de réseaux de citations (http://clair.eecs.umich.edu/aan/index.php) ou pour accroitre les performances de moteurs de recherche spécialisés (http://aclasb.dfki.de/ ou http://saffron.insight-centre.org/acl/).

A l'INRA, l'équipe de recherche Bibliome (unité MaIAGE-INRA) développe de nombreuses applications de TDM appliqué à des textes d'articles scientifiques dans des domaines intéressant l'Inra. Certains de ces projets utilisent également des références (PubMed), des brevets (EspaceNet) et des magazines professionnels (ex. Perspectives Agricoles).

L'UMR Lisis (INRA, ENPC, UPEM) http://www.inra-ifris.org/, associée à l'IFRIS, développe une plateforme Cortext pour l'analyse de corpus textuels (http://www.cortext.net/) dans le cadre de recherches en SHS.

Le Cirad est également concerné au travers de l'UMR IATE (CIRAD, INRA, SupAgro, Université de Montpellier II) développe des méthodes et des outils innovants de traitement de données et de connaissances.http://umr-iate.cirad.fr/axes-de-recherche/ingenierie-des-connaissances/presentation
L'objectif est de proposer des méthodes et outils d'aide à la décision pour le pilotage global de filières de transformation de la biomasse. Ces méthodes et outils doivent permettre de collecter,

représenter et gérer différents types de données et de connaissances, incluant des données imparfaites (par exemple peu fiables, imprécises, ...), des connaissances à dire d'expert et des modèles de génie des procédés. D'autre part, les outils proposés doivent permettre la prise en compte de critères multiples, des préférences et des arguments des acteurs de la filière agricole.

A Irstea, des scientifiques de l'UMR TETIS, utilisent des solutions innovantes de fouille de textes et de données dans le cadre de recherches conduites en collaboration entre IRSTEA, l'INRA et AgroParisTech : fouille d'articles scientifiques pour identifier des thématiques nouvelles de recherche par l'enrichissement de ressources sémantiques (ressources termino-ontologiques de spécialité, thésaurus, etc.) ou améliorer des outils de veille en épidémiologie animale. En outre, le TDM permet la mise en relation de données hétérogènes au sein de corpus très volumineux qui comprennent à la fois des textes scientifiques et non scientifiques, jeux et bases de données, images, etc.), ce qui a permis de découvrir des connaissances nouvelles et complémentaires. Cf. nombreuses références de l'UMR sur des travaux utilisant de texte et data mining (https://tetis.teledetection.fr/index.php/fr/).

L'UMR GESTE (IRSTEA, ENGEES), à Strasbourg (http://geste.engees.eu/), mobilise actuellement des scientifiques qui travaillent sur la problématique des polluants émergents dans l'eau et les changements de pratiques des particuliers et des artisans permettant d'en réduire l'émission. Il s'agit d'une question majeure en termes de santé-environnement, encore insuffisamment cernée sur le plan des enjeux et des solutions. L'usage du TDM permettra de réaliser une cartographie de l'enjeu, dans ses dimensions scientifiques et sociétales, d'en préciser la chronologie, d'identifier des sous-thématiques ainsi que des institutions et acteurs clés.

En matière d'usage du TDM dans le cadre de la conduite de travaux scientifiques, le CNRS, qui présente ses propres arguments par ailleurs, est un acteur majeur. Cf. rapport DIST.

Les enjeux juridiques :

- La demande d'autorisation auprès des éditeurs pour faire du text-mining sur les articles prend énormément de temps et n'est pas toujours couronnée de succès ce qui peut nuire à la conduite des projets scientifiques (corpus incomplets).
- Pour éviter cette étape et conserver un corpus scientifiquement cohérent, les textes traités se limitent souvent à l'exploitation des notices bibliographiques ou aux sources de données gratuites.
- La délégation IST de l'Inra a fait inclure l'usage du TDM dans les licences d'utilisation des plateformes de certains éditeurs (Springer, SGM), mais cela n'est pas satisfaisant car cette autorisation devrait être systématique.

L'urgence : besoin pour toute la recherche française de disposer du droit à faire du TDM sur des corpus mis à disposition dans des formats structurés (XML) avec des licences de type CC-BY

Questions stratégiques, politiques et économiques :

- Les scientifiques français ne sont pas, aujourd'hui, à égalité avec leurs concurrents étrangers, US et européens, Allemagne notamment. En Grande-Bretagne, la refonte du code de la propriété intellectuelle, il y a deux ans, a introduit une exception au droit d'auteur

permettant le TDM à des fins de recherche, sans que cette évolution ne bouscule en rien la bonne santé économique des éditeurs ;

- Des solutions de contournement se mettent en place dans le meilleur des cas : départ vers des laboratoires étrangers non soumis aux mêmes règles et accueil de scientifiques étrangers affiliés à des laboratoires non soumis à la loi française (externalisation du traitement avec risque de perte de la paternité des découvertes pour les laboratoires français concernés) ;
- Un anachronisme en décalage total à l'ère du Big Data et de l'Open Access : les outils TDM existent, sont et seront développés et utilisés par des concurrents (risque majeur de manquer le tournant de la « data science » ;
- Les risques déontologiques sont connus, ils doivent être encadrés par une loi qui permette l'usage sans entrave du TDM pour des activités de recherche, qui favorise l'innovation et qui, de ce fait, ne mette pas sur le même plan la production scientifique et les activités culturelles ;
- Une crainte injustifiée des éditeurs : outre le fait que l'édition scientifique fonctionne sur un modèle différent de celui du secteur culturel, il est important de noter que les corpus utilisés pour des applications de TDM scientifique rassemblent majoritairement des documents pour lesquels les chercheurs ou leurs bibliothèques, en ayant acquis un abonnement pour leurs besoins de consultation documentaire, ont acquis un "droit de lecture". Tous les documents exploités par des outils de TDM ont donc déjà été achetés par leurs usagers : les éditeurs, comme il est normal, ont déjà, en amont des pratiques de TDM, monétisé leurs contenus ;
- Le TDM représente en outre, en soi, un potentiel de retombés économiques pour la France : plusieurs start-up Françaises ont vu le jour suite à des travaux de recherche ayant nécessité le développement d'outils de TDM (au CEA notamment) en lien avec des partenaires privés. Aider l'usage du TDM, c'est aussi favoriser l'économie française du numérique.