Du texte à la connaissance : fouille de texte, analyse textuelle, enrichissement sémantique de contenus, annotation sémantique…

Mader C.

Cahier des Techniques de l'INRA, Numéro spécial 2012
pp. 52-60

Cet article est tiré du numéro spécial 2012 du Cahier des Techniques de l'INRA :

Marchoux E. (Coord.) & Hologne O. (Dir.) **L'Information Scientifique et Technique à l'Inra, des compétences au service de la recherche. Retour d'expérience sur des projets, services, outils et méthodes.** *Cahier des Techniques de l'INRA* Numéro spécial 2012. Paris : Inra, 2012. 141 p.

Cahier des Techniques de l'INRA, ISSN 0762 7339
Institut National de la Recherche Agronomique - 147 rue de l'Université - F-75338 Paris Cedex 07
https://intranet.inra.fr/cahier_des_techniques

Du texte à la connaissance : fouille de texte, analyse textuelle, enrichissement sémantique de contenus, annotation sémantique…

Claudine Mader[1]

Résumé. « L'étiquette « Extraction de Texte » (Text Mining) a été utilisée pendant un certain nombre d'années comme un terme générique pour se référer aux approches qui abordent ces questions. » (Mayer, 2008). À l'heure actuelle, on parlera plutôt « d'analyse de contenus » plus proche de la réalité des faits. Dans le cas d'applications nécessitant le traitement de milliers de ressources de formats divers, le traitement automatique du langage « permet de mettre en place une automatisation du processus d'acquisition de la connaissance et l'utilisation de cette connaissance acquise pour l'annotation des ressources utilisées » (Armadheil, 2007).

« Il existe une quantité phénoménale de pages écrites en langage naturel et contenant de nombreuses connaissances exprimées dans des langues très diverses. Ces connaissances, bien que présentes, ne sont pas présentées de manière structurée. Des procédés sont déjà utilisés pour repérer de manière automatique la ou les thématiques d'un texte. Le marquage de nom de personnes ou de lieux, appelée extraction d'entités nommées, en est un exemple. Il est également possible d'y adjoindre des techniques plus proches des considérations de l'analyse sémantique afin de déterminer plus précisément la thématique et même de parvenir à extraire du sens » (Poilbeau, 2010). Après avoir brièvement décrit les technologies et le contexte dans lequel elles ont été utilisées, nous essaierons de montrer, à partir de quelques exemples qui les illustrent, ce qu'elles apportent comme valeur ajoutée.

Mots clés : analyse de contenus, annotation sémantique, extraction d'information, text mining, gestion des connaissances, terminologies métiers, ontologies

Contexte

Traditionnellement dans le monde de la publication scientifique, l'unité considérée est toujours le document. La plupart de l'infrastructure aussi bien que les habitudes de nombreux acteurs dans ce domaine sont centrées autour de la notion de documents : c'est le document qui est publié, stocké, vendu, recherché et téléchargé.

« Il peut être discuté cependant que dans la plupart des cas le document comme tel ne représente pas la granularité idéale : un chercheur cherchant des indices de cibles potentielles dans le développement d'un médicament ou des façons d'améliorer l'efficacité d'un algorithme dans l'informatique ne peut accepter le fait que pour parvenir à la réponse exigée, il devrait rechercher, télécharger, imprimer et lire des documents ; en fin de compte dans beaucoup de ces scénarios le document est juste la forme traditionnelle de livraison, le conteneur pour les pièces d'informations dans lesquels le chercheur trouvera sa solution » (Mayer, 2008).

Contribuer à la mise en place d'outils et de méthodes pour gérer les connaissances fait partie des missions de la Direction de la Valorisation/Information Scientifique et Technique (DV-IST) car ces technologies peuvent apporter une réponse aux recommandations du Contrat d'objectifs Inra 2012-2016 ; à l'heure des grandes Alliances et des Métaprogrammes, pour promouvoir les innovations, il faut « mobiliser les ressources autour des priorités et des défis scientifiques ». « L'innovation à l'Inra est la résultante d'un large processus interactif avec les acteurs socio-économiques au sein de réseaux multipartenaires où **la place de l'ingénierie des connaissances, en particulier en raison des dimensions intégratives de la démarche et de la pression croissante des attentes sociétales, mérite d'être clarifiée et renforcée** ».

Dans le cadre d'une convention de partenariat de Recherche & Développement (décembre 2010 à juin 2012) entre la société Témis (Société française spécialisée en annotation sémantique de contenus) et la DV-IST, ces technologies d'annotation sémantique ont pu être testées, l'objectif étant de créer des applications en collaboration avec

1 INRA, UAR 1266, DV-IST Pôle Gestion des connaissances, F-78026 Versailles Cedex, France ; Claudine.Mader@versailles.inra.fr

les experts fonctionnels de Témis, les ingénieurs de la connaissance de la DV-IST et l'aide d'experts scientifiques, pour arriver à une preuve de concept[2].

La plateforme Luxid[3] de Témis utilise des « cartouches de connaissances » (*skill cartridges*) qui décrivent le vocabulaire conceptuel dans lequel les règles métiers s'expriment. Ces cartouches couvrent partiellement les domaines de l'alimentation, de l'agriculture et de l'environnement ; un des objectifs de la coopération Témis/DV-IST est de construire une cartouche Agronomie de façon à mettre à jour le modèle d'annotation, donc d'ajouter de la pertinence à l'analyse sémantique des contenus.

Quelques définitions

Avant d'aborder le sujet, il est nécessaire de clarifier le domaine par quelques définitions.

Sémantique : dans le cas de la gestion des connaissances, il s'agit de la formalisation du sens des contenus (description du sens des pages dans un langage compréhensible par une machine) permet par exemple de les rendre accessibles à des processus automatisés.

L'annotation sémantique : elle a pour objectif d'exprimer la « sémantique » du contenu d'une ressource afin d'en améliorer sa compréhension, sa recherche et donc sa réutilisation par les utilisateurs finaux. (Armadeilh, 2007).

Métadonnées d'annotation documentaire : le Dublin Core fait office de standard pour l'annotation avec des descripteurs tels que l'auteur, le titre, la source, l'éditeur, la date de publication, le sujet, etc.

Technologies d'annotation sémantique : il s'agit de technologies avancées permettant de transformer du texte libre en une série de données analysables. « Ces méthodes substituent les **concepts** eux mêmes aux **mots clés** comme composants principaux des métadonnées. En introduisant cette dimension conceptuelle dans des index auparavant plats, l'enrichissement de contenu sémantique a également changé l'utilisation qui peut en être faite » (Mayer, 2011).

Ontologie : une ontologie est une spécification formelle, explicite et consensuelle de la conceptualisation d'un domaine (Gruber, 1993). Une ontologie est constituée d'un ensemble de concepts organisés hiérarchiquement et structurés par des rôles liant ces concepts (Ma *et al.*, 2009). Dans le processus d'annotation sémantique les ontologies jouent un rôle essentiel car elles offrent une modélisation des concepts et expriment de façon formelle les attributs et les relations qui vont servir à annoter les contenus des ressources.

Objectifs et descriptif du projet

La DV-IST souhaite préparer le déploiement d'une solution de ce type adaptée aux thématiques de l'Inra ; pour cela elle a été accompagnée par la société Témis dans un projet qui vise, à partir de traitement de données brevets, de notices bibliographiques, de littérature scientifique en texte intégral et de bulletins de veille, à établir un démonstrateur[4] sur les focus suivants : biotechnologies vertes, nutrition santé, OGM maïs et coton, grandes cultures.

Les trois partenaires de la convention de recherche sont :
- la société Témis productrice de la plateforme Luxid ;
- le pôle Gestion des connaissances (GeCo) de la DV-IST ;
- les experts scientifiques : Inra Transfert pour les deux premières thématiques, un expert spécialisé pour les OGM et le GIS GC HP2E[5] pour les grandes cultures.

Les résultats de ce partenariat doivent permettre d'enrichir les « cartouches de connaissances » Témis dans le domaine de l'agronomie et de la biologie végétale.

La DV-IST attend de cette expérience la validation des bénéfices obtenus par ces technologies du point de vue de l'accès à l'information fournie.

2 *Démonstration de faisabilité venant de l'expression anglaise Proof of concept (POC).*
3 *La plateforme Luxid est une solution logicielle d'enrichissement sémantique de contenus.*
4 *Synonyme de « preuve de concept » (voir [2]).*
5 *GIS GC HP2E : groupement d'intérêt scientifique grandes cultures à hautes performances économiques et environnementales.*

Quatre thématiques ont été retenues

Biotechnologies vertes

Les grandes cultures (dont maïs, colza, blé) et cultures maraîchères (dont tomate, concombre, etc., pour les brevets uniquement). Les sous-thèmes associés :

- les résistances à la sécheresse, aux agents pathogènes, aux maladies ;
- l'adaptation de ces variétés aux changements de conditions climatiques ;
- l'utilisation de l'azote.

Sources utilisées : 13 940 notices WOS[6] et 18 729 notices CAB Abstracts[7] dédoublonnées avant leur entrée dans Luxid, 11 185 brevets QPAT[8] (résumé, revendications, description), des bulletins de veille sur la phytoprotection et les semences.

Nutrition-Santé humaine

L'identification de nouveaux alicaments pour la prévention de maladies en lien avec les allégations[9] santé. Il peut s'agir de nouveaux aliments, compléments alimentaires, ingrédients et composés ; sont également suivis les comportements alimentaires.

Sources utilisées : 13 743 notices WOS et 2 783 notices FSTA dédoublonnées avant leur entrée dans Luxid.

OGM

La base de données de trois experts (Jean Baptiste Bergé, Agnès Ricroch et Marcel Kuntz) spécialisés en transgénèse végétale pour le maïs et un corpus constitué par la DV-IST pour le coton.

Sources utilisées : références annotées pour la base maïs et notices de la base CAB Abstracts pour le coton.

Grandes cultures à hautes performances économiques et environnementales

Le focus choisi pour le prototype est l'implantation des cultures.

Sources utilisées : le corpus est constitué de 19 012 notices de la base CAB Abstracts et 400 documents de littérature grise en texte intégral (brochures, rapports, livrables de projet (CASDAR, ANR, européens), etc.), ainsi que des articles scientifiques en texte intégral.

Les enjeux du projet

L'étude préalable doit tester la démarche sur un nombre limité de documents représentatif de la base de connaissance envisagée afin d'apprécier la valeur ajoutée qu'une telle démarche pourrait apporter aux membres du GIS.

Fédérer différents corpus de formats variés : les projets pilotes ont utilisé des sources provenant de bases de brevets, de bases de données bibliographiques scientifiques, des bulletins de veille, des documents en texte intégral de formats très différents (xml, Word, Excel, pdf, ppt).

Extraire les informations spécifiques concernant les domaines de recherche de l'Inra (Agriculture – Alimentation – Environnement). Les entités biologiques mais aussi économiques sont à traiter. L'Inra dispose d'un certain nombre de vocabulaires spécifiques couvrant les domaines qu'il traite (à différents niveaux d'organisation : listes à plat, taxonomies, ontologies) ; les possibilités et les modalités d'intégration de ces ressources dans la plateforme Luxid sont à évaluer.

Mettre en valeur dans les textes des éléments porteurs d'information : entités (biologiques, chimiques, économiques) et termes techniques dans un objectif d'extraction de connaissances et d'aide à la lecture de gros volumes d'information.

Organiser les sources en fonction de leurs similarités et singularités pour l'aide à l'analyse de gros corpus par des méthodes de catégorisation et de classification (thématique, géographique, structurelle, temporelle). Le retour aux données doit être possible pour permettre un affinage ou une correction des résultats d'analyse.

6 *WOS : Web of Science, base de citations produite par Thomson Reuters.*
7 *CAB Abstracts : base de données bibliographique sur les domaines des sciences de la vie appliquées à l'agriculture, l'environnement, les sciences vétérinaires, l'économie appliquée, l'alimentation et la nutrition produite par les Commonwealth Agricultural Bureau.*
8 *QPAT : base de données de brevets.*
9 *Tout message ou représentation, non prescrit par la législation communautaire ou nationale, qui affirme, suggère ou implique une relation entre une denrée alimentaire et la santé.*

Adapter des ressources sémantiques aux besoins de la DV-IST qui souhaite mettre en place un serveur terminologique au service des chercheurs de l'Institut.

Analyser et visualiser l'évolution temporelle et spatiale des données.

Construire et visualiser des réseaux (d'acteurs ou d'autres objets connectés).

Méthodologie

La plateforme Luxid utilise différents modules d'extraction sémantique :

- des **ressources linguistico-sémantiques** : thesauri, taxonomies, ontologies ;
- des **systèmes à base de règles** : expressions régulières (modèles permettant de manipuler des chaînes de caractères), raisonnements morpho-syntaxiques (lemmatisation, affectation de catégories grammaticales, identification des mots et de groupes nominaux (entités nommées et termes)) et l'exploitation de synonymes, la recherche d'expressions, l'extraction de connaissances basées sur des dictionnaires paramétrables ou des ontologies ;
- des **mécanismes statistiques** : mécanismes d'apprentissages, extractions pertinentes pour un concept donné qui permettront d'organiser les textes analysés par thèmes et par classes.

L'extraction des connaissances et l'analyse utilisent ces technologies avancées du Text-Mining pour transformer du texte libre en une série de données rapidement analysables par un utilisateur.

Résultats

Une approche visuelle interactive

La présentation des résultats de ces projets consistera à restituer en images la réponse à quelques questions que l'on peut se poser sur ces corpus. Certaines de ces réponses sont proposées en standard par la plateforme, d'autres sont à élaborer au gré des besoins de l'utilisateur final.

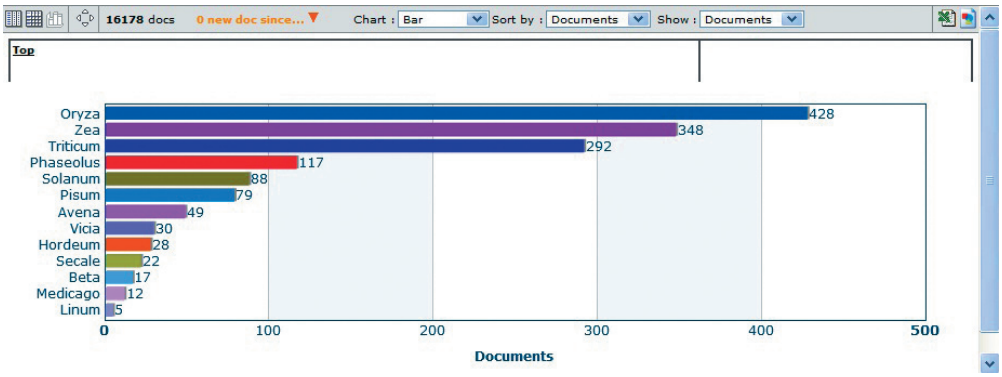

Figure 1. Quelle répartition des plantes de grande culture dans les publications ?

On peut choisir un élément du graphique pour affiner ses résultats et commencer d'autres analyses. Tous les éléments du graphique sont cliquables et l'on peut toujours passer d'une vision graphique à une vision textuelle des résultats.

À partir du choix d'une plante précise dans la Figure 1 (Zea), le système peut analyser une question spécifique.

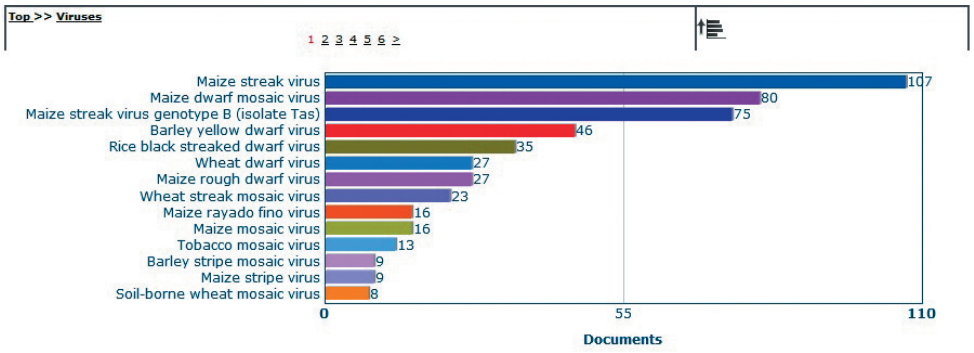

Figure 2. *Quels sont les virus phytopathogènes sur le maïs par ordre d'importance dans le corpus considéré ?*

L'analyse peut porter sur les métadonnées Dublin Core des documents comme les pays de publication (Figure 3) (paramétrables au niveau mondial, continental ou d'un groupe de pays précisé) ou sur les auteurs publiant sur le sujet (Figure 4).

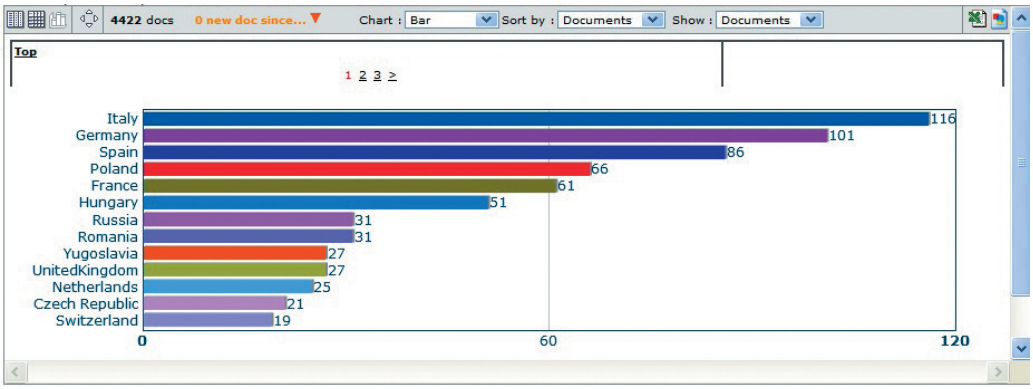

Figure 3. *Quel est le positionnement des différents pays d'Europe publiant sur le maïs ?*

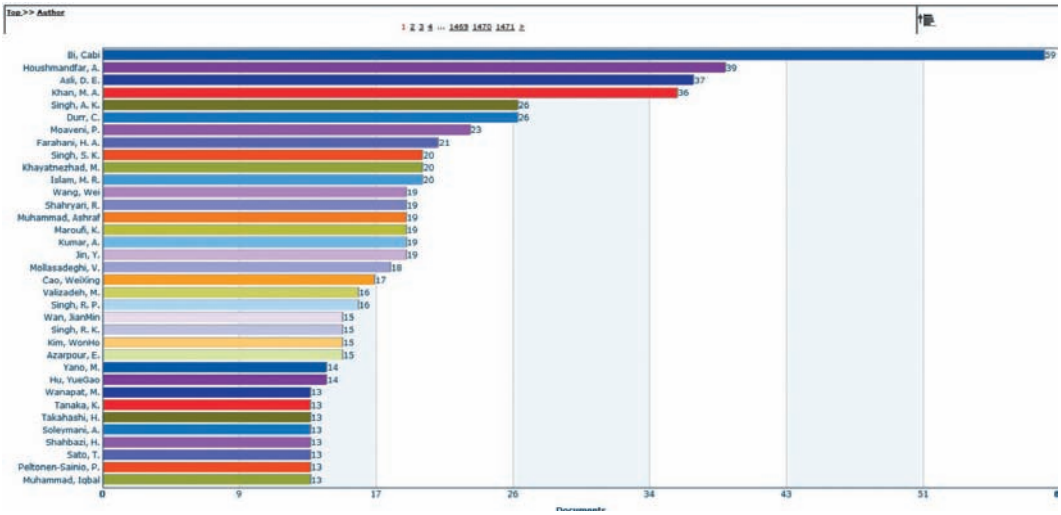

Figure 4. *Qui publie sur ce sujet ?*

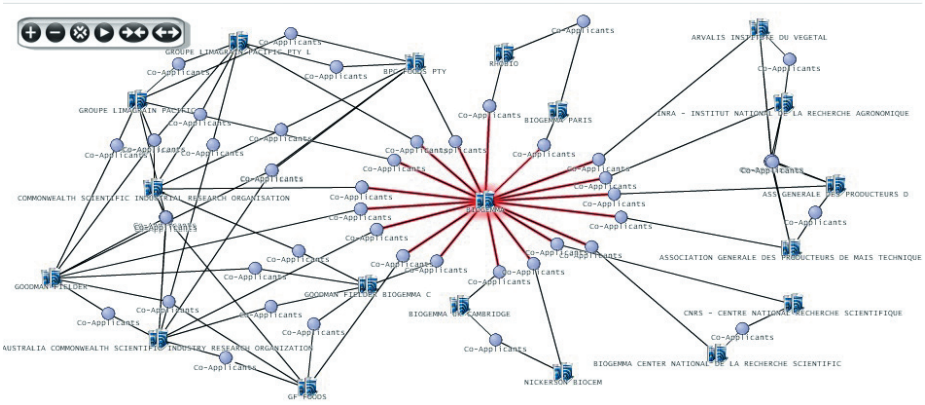

Figure 5. Réseau de Biogemma en dépôt de brevets sur la génétique du blé.

Une approche contextuelle et connaissances

En fonction des annotations apportées par ce type de plateforme d'analyse de contenu au moyen des « cartouches de connaissances » déjà établies ou enrichies lors des projets, l'information est contextualisée : la lecture par concepts est très rapide, elle permet d'arriver à l'information pertinente en quelques clics.

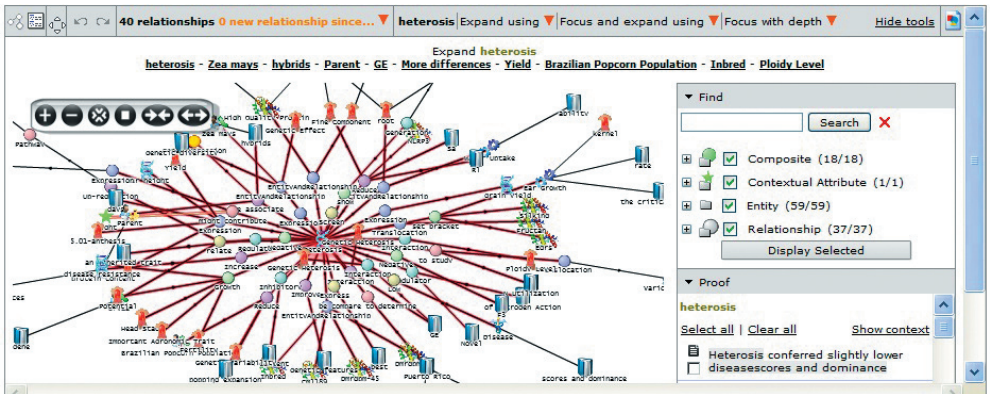

Figure 6. Quels concepts sont liés à l'hétérosis chez le maïs, quels types de relation entre ces concepts ?

Des réseaux de proximité entre concepts peuvent être calculés automatiquement (Figure 7).

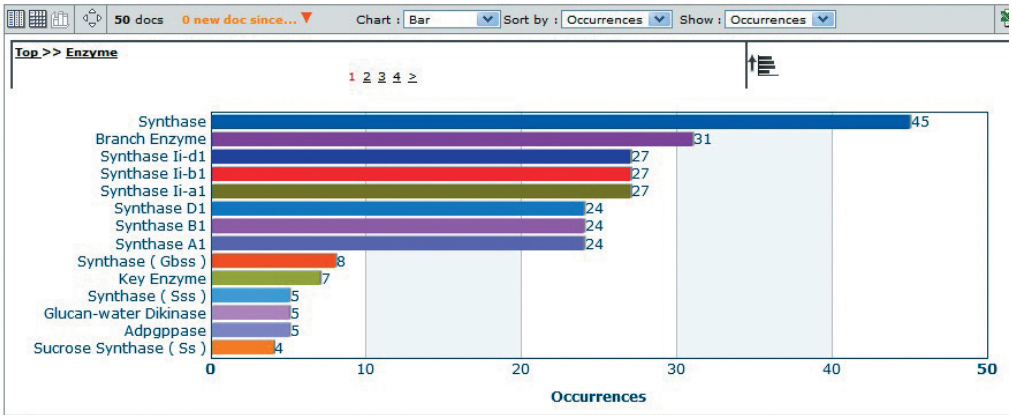

Figure 7. Quelles enzymes interviennent dans la synthèse de l'amidon ?

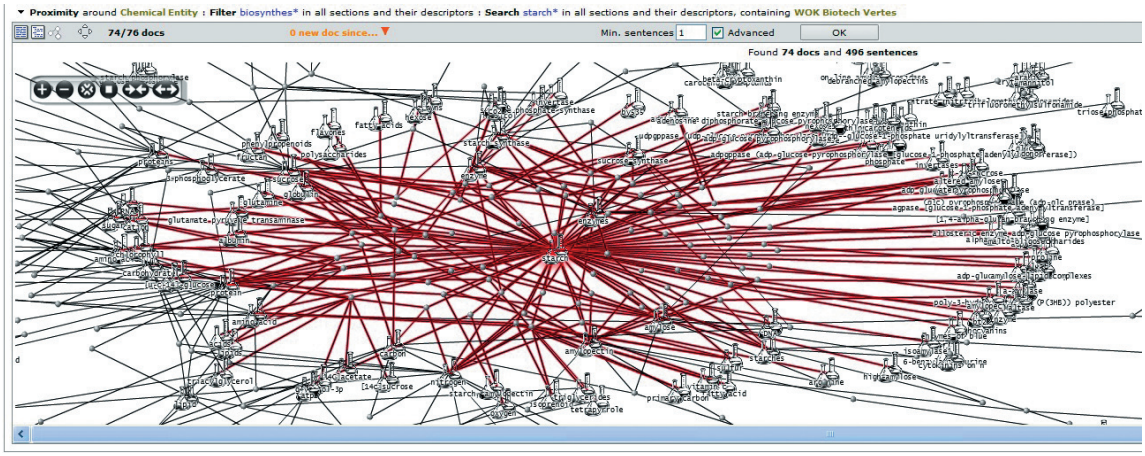

Figure 8. Quels éléments interviennent dans la biosynthèse de l'amidon ?

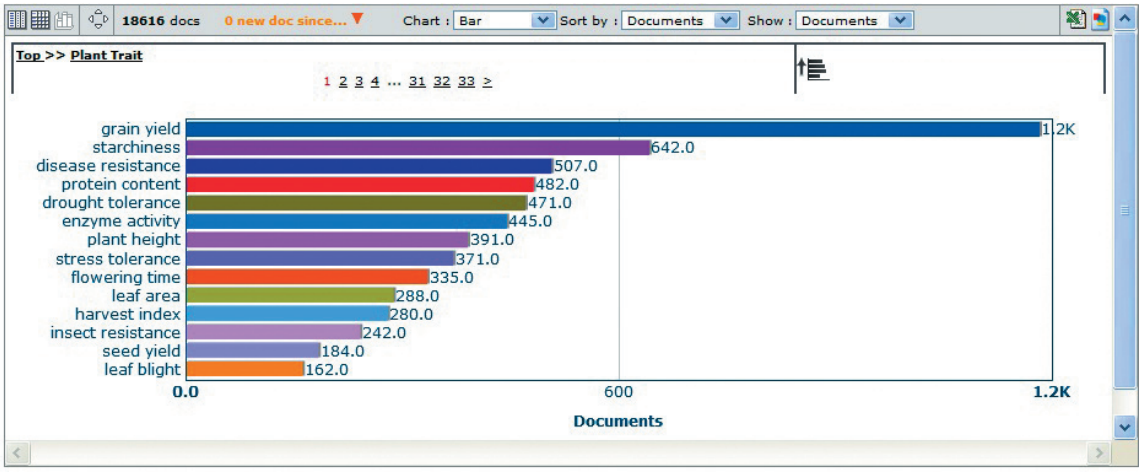

Figure 9. Analyse par traits fonctionnels dans le corpus constitué sur le maïs.

On peut également aborder le feuilletage des connaissances contenues dans les textes. Le surlignage des éléments appartenant à des entités thématiques différentes cible rapidement le contenu d'un document.

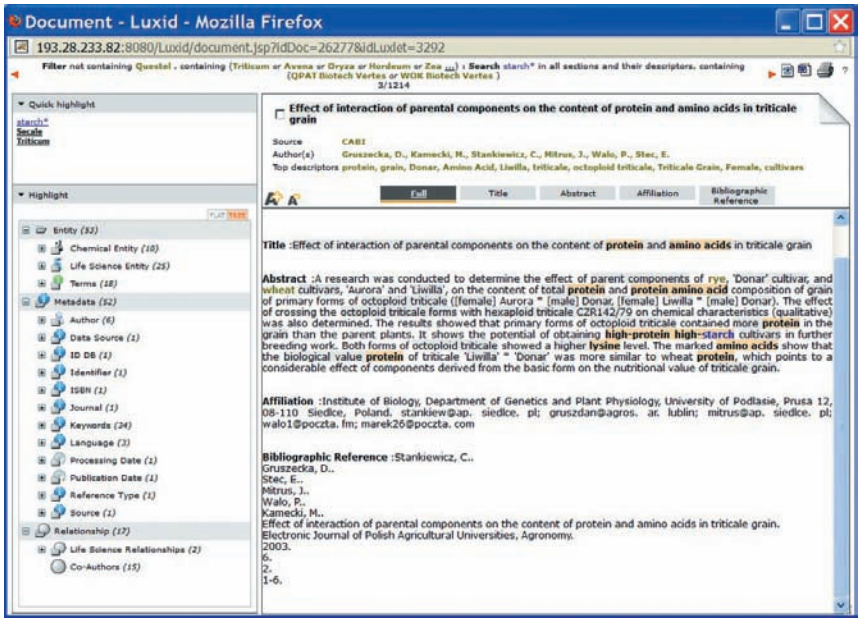

Figure 10. Quelles sont les entités chimiques étudiées dans ce texte ?

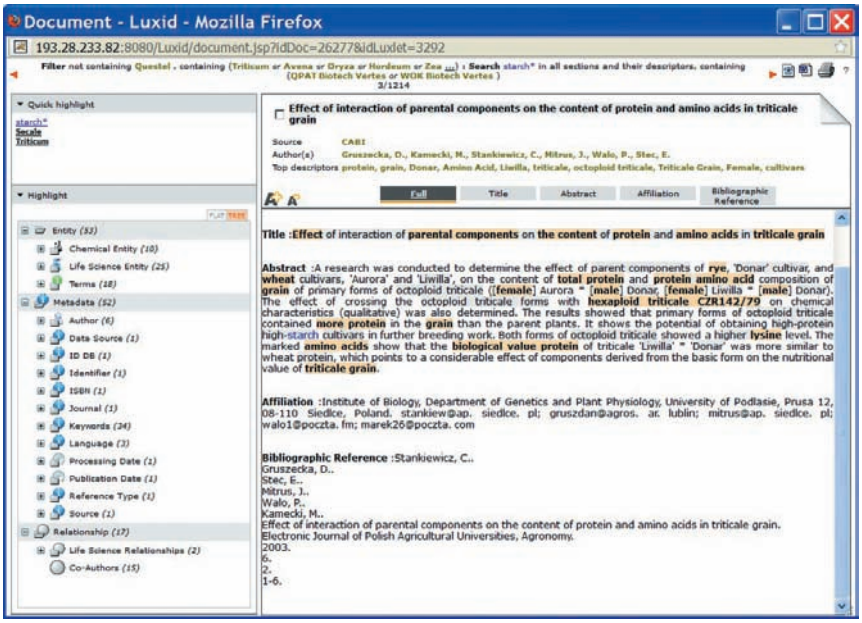

Figure 11. Pour le même texte, quelles sont les entités biologiques ?

Le feuilletage des connaissances contenues dans le texte répond à des questions plus complexes (Figure 12).

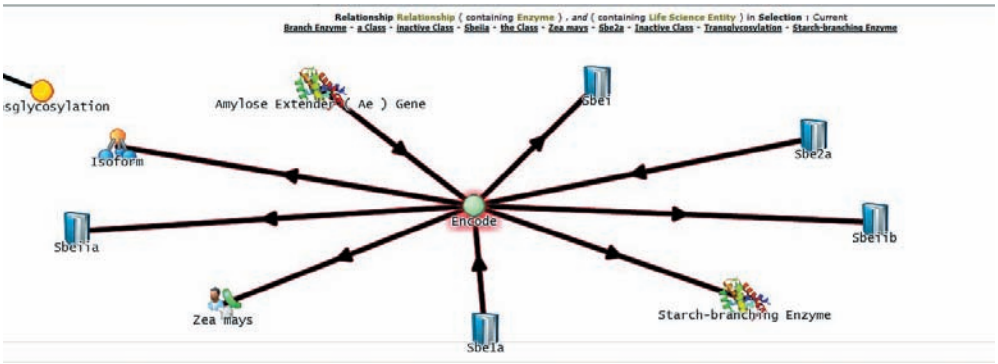

Figure 12. Chez le maïs quels sont les 3 isoformes de la starch branching enzyme qui sont encodés par les gènes Sbe1a, Sbe2a, and Amylose extender (Ae) ?

Tous ces éléments peuvent être exploités dans des tableaux de bord thématiques regroupant les informations que l'on choisit de suivre et permettent de surveiller un sujet en se mettant à jour automatiquement dès l'entrée de nouvelles données (Figure13).

Figure 13. Tableaux de bord.

Discussion et conclusion

Les technologies d'analyse de contenu apportent une forte valeur ajoutée à la gestion des connaissances dans le monde de la recherche scientifique. Elles ont acquis suffisamment de maturité pour offrir dans une même application :

- une aide à la lecture de gros volumes d'information ;
- une analyse « en contexte » de la connaissance ;
- des réseaux de concepts ;
- une surveillance de thématiques ciblées, de fronts de science, des concurrents, des collaborateurs potentiels, des brevets déposés dans son domaine, de l'évolution d'une thématique dans le temps, des trous de connaissance, etc. ;
- une valeur ajoutée importante dans la lecture non plus par mots-clés mais par concepts avec mise en contexte ;
- l'intégration d'ontologies dans ces systèmes n'a pas été testée dans le cadre de cette convention car c'est la V.6 de Luxid, parue cette année qui permet d'intégrer des ontologies.

Le secteur des Sciences de la Vie a pris rapidement conscience des défis liés à la croissance des flux d'information et cherché des solutions. Le Text Mining associé à des fonctions sémantiques d'analyse de contenus en est une prometteuse.

What Else ?

Si cet article vous a intéressé, contactez-nous : geco@versailles.inra.fr

Références bibliographiques

Armadeilh F (2007) Web sémantique et informatique linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle, Thèse de Doctorat, mai 2007.

Gruber TR (1993) A translation approach to portable ontologies, *Knowledge Acquisition*, 5, 199-200.

Poilbeau T (2010) Du TAL au web sémantique, Conférence Du TAL au Web sémantique, mars 2010.

Mayer D (2008) New methods to access scientific content, Information Services & Uses, 28,141-146.

Mayer D (2011) Introduction à l'enrichissement sémantique de contenu, Livre blanc Témis S.A., septembre 2011.

Ma Y, Audibert L, Nazarenko A (2009) Ontologies étendues pour l'annotation sémantique, MaEtAl_IC2009_57.