INTÉRÊT ET MÉTHODE D'EXTRACTION DE L'INFORMATION SCIENTIFIQUE CHINOISE

Nadège GUÉNEC
Doctorante, Université Paris-Est Marne la Vallée
Membre du Groupe de Recherche en Intelligence Économique, Centre National de la Recherche Scientifique (CNRS)
Henri DOU
Professeur, École Supérieure de Commerce et de Management (ESCEM)
Membre du Groupe d'Expert en Intelligence Économique, Centre National de la Recherche Scientifique (CNRS)

▪ L'industrie chinoise de l'information s'est déployée de concert avec l'émergence d'Internet et le développement impressionnant de la recherche scientifique chinoise. Nous présentons dans cet article le portail de ressources bibliographiques chinois le plus important car développé par le gouvernement chinois à l'université de Qinghua à Pékin. La maîtrise de la recherche documentaire et de l'analyse des ressources disponibles sur ce portail nous semble indispensable pour tout veilleur scientifique dans les domaines de pointe. Ceci souligne la nécessité d'une allocation de ressources entre autres au niveau de la recherche et de l'analyse des informations chinoises. Cette étude tente de démontrer grâce à un exemple comparatif d'une recherche sur un même thème sur des bases d'articles scientifiques américaines et chinoises le fait que le monde de l'information scientifique et technique est en train de changer et que "l'universalité" des sources d'information anglo-saxonnes est actuellement remise en question.

▪ De Chinese informatie-industrie heeft zich ontvouwen "in goed overleg" met het verschijnen ven het internet en de impressionante ontwikkeling van het Chinees wetenschappelijk onderzoek. Wij stellen in dit artikel dé omvangrijkste – want ontwikkeld door de Chinese regering aan de Tsinghua universiteit te Beijing – Chinese bibliografische *portal* voor. De beheersing van het documentaire opzoeken en de analyse van beschikbare (hulp)bronnen op deze *portal* lijkt ons onontbeerlijk voor alle wetenschappelijke informatiebewakers in de speerpuntdomeinen. Dit onderstreept de noodzakelijkheid van het toekennen van middelen onder andere op het niveau van het onderzoek en de analyse van Chinese informatie. Deze studie probeert aan te tonen dankzij een vergelijkend voorbeeld van een opzoeking over eenzelfde onderwerp in Amerikaanse en Chinese artikelendatabanken dat de wereld van de wetenschappelijke en technische informatie aan het veranderen is en dat de "universaliteit" van de Angelsaksische informatiebronnen tegenwoordig in vraag gesteld wordt.

Ce travail s'inscrit dans le cadre du développement de systèmes d'information en intelligence économique mettant en évidence la nécessité de développer des politiques nouvelles d'acquisition des informations stratégiques, compte tenu des nécessités de la mondialisation. Il vise à une utilisation des outils de traitement automatique de l'information développés par la recherche publique française pour une application à des "territoires" nouveaux de l'information scientifique et technique mondiale, parmi lesquels l'information chinoise occupe une place d'importance croissante.

Au cours de la deuxième moitié du XXe siècle, le développement de l'informatique et des techniques de communication a favorisé le passage à l'ère de l'information. Une nouvelle industrie, celle du savoir, est née. Celle-ci est essentiellement portée par les bases de données qui sont les réceptacles du savoir humain dans les divers domaines de connaissance.

La Chine, s'ouvrant au marché international au début des années 1980 a dû s'adapter à cette industrie afin de s'enraciner selon son souhait, dans le commerce international et devenir compétitive sur le marché mondial. Le développement fulgurant d'Internet en Chine, à partir de la seconde moitié des années 1990, permet l'émergence du marché de l'information qui, en Chine comme ailleurs, va devoir se structurer[1]. L'industrie des bases de données est un des aspects les plus importants du secteur de l'information scientifique et technique (IST)et peut être utilisé comme indicateur pour évaluer le degré d'appartenance d'un pays à un système économique dans lequel l'information a acquis une valeur marchande dominante.

Nous proposons dans cet article la présentation d'un portail regroupant plusieurs bases de données: le portail *CNKI*[2], China National Knowledge Information.

La lecture de cet article intéressera tout chercheur désireux de se tenir au courant des évolutions de la recherche scientifique mondiale dans son domaine d'application.

Nous précisons qu'ayant réalisé ces travaux dans le cadre d'un partenariat industrie-université, des clauses de confidentialités ne nous permettent pas d'exposer dans cet article les résultats

approfondis de nos recherches. Nous nous focalisons ici sur les méthodes de recherche et notamment sur la présentation des sources d'information scientifique chinoise et non sur le fruit de ces recherches. Cependant, nous informons le lecteur que les résultats des recherches effectués à partir des sources présentées ici ont été validés dans leur pertinence par les chercheurs industriels du secteur et à qui sont destinés les connaissances produites par nos travaux.

Le Web chinois, une économie grandissante

Selon le rapport annuel sur le développement d'Internet en Chine publié en janvier 2008 par le CNNIC (China Internet Network Information Center)[3], organisme d'État, en décembre 2007, le nombre d'internautes chinois avait dépassé les 210 millions (voir figure 1), faisant de la Chine le 2e pays utilisateur d'Internet au monde après les États-Unis (270 millions). Affichant une progression du nombre d'utilisateurs de plus de 53 % sur l'année 2007, il y a fort à parier que la Chine deviendra le 1er utilisateur mondial d'Internet dans le courant de l'année 2008.

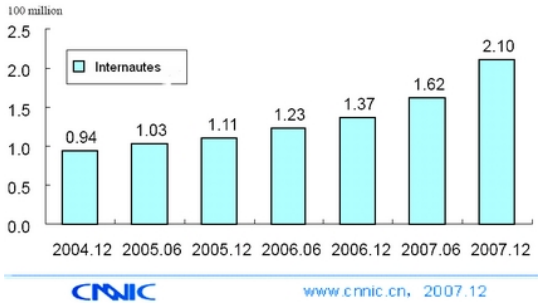

Fig. 1 : Croissance du nombre d'internautes en Chine.

Au niveau mondial, le nombre de noms de domaine et de pages Web a connu une croissance de plus de 60 % durant l'année 2007 et le nombre d'adresses IP s'élève à 135 millions, soit une augmentation de 38 % pour l'année. Le Web chinois comptabilisait, au 31 décembre 2007, 1,5 million de sites Internet.

En ce qui concerne l'industrie des bases de données, la Chine fait partie du Comité International des Bases de Données Scientifiques et Techniques depuis 1984. En 1987, elle s'est dotée d'un centre national d'information qui avait alors décidé de la création de 134 grandes bases de données. Moins de dix ans plus tard, elle en comptabilisait plus de 1300. À l'heure actuelle, près de la moitié des bases de données sont développées par des organismes publics et sont consultables sur le réseau Internet.

Tab. 1: Croissance des ressources Internet fondamentales par quantité

	Total as of December 2007	Qty per 10,000 persons	Qty per 10,000 netizens	Total as of December 2006	Annual Growth Rate
IPv4 address (Nrs)	135,274,752	1,029	6,442	98,015,744	38.0%
domain name(Nrs)	11,931,277	91	568	4,109,020	190.4%
Including: CN domain name (Nrs)	9,001,993	68	429	1,803,393	399.2%
Website (Nrs)	1,503,800	11	72	843,000	78.4%
Web pages (Nrs)	8,471,084,566	64,444	403,385	4,472,577,939	89.4%
Web bytes (KB)	198,348,224,198	1,508,948	9,445,154	122,305,737,000	62.2%
international outlet bandwidth(Mbps)	368,927	3	18	256,696	43.7%

Les données proviennent du rapport du CNNIC de janvier 2008.

Intérêt de l'IST chinoise pour un chercheur occidental

Selon un rapport de l'OCDE[4], la Chine est devenue le numéro deux mondial de la dépense en recherche et développement, injectant 136 milliards de dollars américains en 2006, soit une hausse de plus de 20 % sur un an. (USA : 330 milliards de dollars pour la même période). Le journal *Nature* affirme[5] que le développement scientifique de la Chine se fait à un rythme encore plus fort que celui du développement économique car il doit servir ce dernier. Même si l'économie s'est fortement libéralisée, les lignes directrices et les axes stratégiques sont déterminés par l'État, notamment par le biais des plans quinquennaux. Le gouvernement chinois, conscient du clivage social, intellectuel, culturel de sa population et désireux de se positionner comme un État fort sur le plan international, a émis des stratégies très précises en matière de recherche. Il se dégage de la lecture de ces directives la volonté claire du gouvernement chinois de positionner le pays parmi les leaders internationaux sur un certain nombre de technologies clefs. Le président Hu Jintao, a d'ailleurs déclaré en 2006 que "*en 2020, la Chine sera l'un des pays les plus innovateurs de la planète*". Pour cela, le gouvernement affiche un taux de dépenses publiques pour la science et la technologie en augmentation constante d'environ 20 % par an depuis 1995.

La stratégie est de subventionner fortement certains domaines de recherche fondamentale et expérimentale et de protéger le savoir produit au moyen de la prise de brevets afin de donner un avantage compétitif au pays. Les secteurs visés par ces programmes (notamment le programme 863) sont les technologies de l'information et de la communication, les nanotechnologies, les biotechnologies, l'optique, l'énergie,

l'environnement et le développement durable, la pharmacie, la sécurité alimentaire.

On a pu constater, durant l'année 2007, une forte explosion des publications scientifiques chinoises du plus haut niveau international dans tous les secteurs de la recherche. Pour illustrer ce propos, nous proposons en annexe A et B une requête identique lancée une première fois en anglais et une seconde fois en chinois sur le moteur *Google Scholar*. *Google Scholar* n'est pas le plus performant des moteurs spécialisés dans la recherche de publications scientifiques mais il est aujourd'hui encore le seul à posséder une interface en langue chinoise. Dans sa version occidentale, *Google Scholar* indexe les documents de nature académique des sites Web de certains éditeurs commerciaux, sociétés savantes et universités. Dans sa version chinoise, *Google Scholar* référence l'intégralité de la base *CqVip*, le plus gros éditeur chinois de publications scientifiques privé, concurrent direct de la base publique *CNKI* que nous présenterons ci-après.

Les résultats de notre expérience permet de constater qu'une requête lancée en anglais sur *Google Scholar* fait apparaître en auteurs les plus cités des chercheurs chinois et où la même requête, effectuée en caractères chinois liste des auteurs différents de ceux apparaissant en premier dans la requête en anglais. Cela mets d'emblée en évidence un potentiel de chercheurs chinois plus important que celui que l'on pourrait envisager à première vue, la majorité d'entre ceux-ci n'apparaissant pas dans le Web occidental.

De plus, nous constatons, après une recherche rapide sur les bases de données internationales (ici *Current Contents*) et le Web chinois (recherche classique sur *Google Scholar* et *Baidu* en utilisant des caractères chinois) que les articles chinois proposés en langue anglaise ne reflètent qu'une partie émergente de la recherche chinoise.

La figure 2 présente une carte de la recherche chinoise sur le sujet du blé hybride réalisée fin 2006 à l'aide de la base *Current Contents* (langue anglaise).

Fig. 2 : Principales provinces concernées selon Current Contents.

À l'aide simplement de nos connaissances sur la Chine et d'une exploration classique du Web chinois (par moteur de recherche, *Google* et *Baidu*), nous avons pu établir sur la figure 3 le portrait de la Chine dans le domaine du blé hybride, à la fin de l'année 2006.

Nous constatons donc, à la vue de ces deux cartes, que bien des laboratoires de recherches chinois ne sont pas apparent sur le Web occidental. Cela vient essentiellement du fait qu'un nombre restreint d'articles scientifiques chinois sont publiés dans les revues occidentales (américaines la plupart du temps). Les chercheurs chinois publient avant tout dans les publications nationales chinoises. Étant donné la croissance de la recherche chinoise et son excellence dans bien des domaines scientifiques, il devient donc essentiel, pour le veilleur scientifique occidental, de pouvoir mener des actions de veille et de recherche sur les publications scientifiques chinoises, faute de quoi, il risque d'être pris au dépourvu lors de la publication, forcément a posteriori, du même document en langue anglaise.

Fig. 3: Principales provinces concernées selon Baidu.

La même remarque, mais à un niveau supérieur peut être faite en ce qui concerne les décideurs dans le domaine de la recherche en France. La connaissance des travaux mondiaux ne doit pas se limiter simplement à l'utilisation des bases de données américaines. De plus en plus de travaux de qualité sont publiés dans des langues orientales ou dans des langues nationales différentes

de l'anglais. En outre cela pose aussi le problème de l'équipement des centres de documentation et des laboratoires de recherche de pointe en "équipements logiciels" de traduction des langues orientales (entre autres le chinois et le japonais). Si le système de traduction Systran est d'origine française et est considéré comme un des meilleurs au monde, force est de constater qu'il n'est que peu ou pas utilisé en France au niveau de la recherche.

Présentation de *CNKI*

Nous espérons avoir convaincu le lecteur de l'intérêt d'effectuer des recherches d'information et des veilles automatisées sur le Web chinois et nous proposons à présent de nous focaliser sur le portail d'information *CNKI*. L'industrie chinoise des bases de données et des diverses ressources de publications électroniques connaît un essor rapide depuis le début des années 2000. Il va de pair avec la croissance générale d'Internet. On trouve sur le marché privé quelques grands éditeurs de contenus académiques intéressants comme *Wanfang Data* ou *CqVip*. Ces fournisseurs de bases de données proposent des périodiques, des thèses et des contenus scientifiques d'autres natures (comme des actes de colloques par exemple)[7,8]. La base dont nous allons faire la présentation pratique est un projet universitaire, donc public et est fortement soutenu par le gouvernement chinois qui compte s'appuyer sur ce portail pour stimuler la croissance de la société de l'information chinoise ou le développement en Chine de l'intelligence par l'information.

CNKI est le sigle pour "*China National Knowledge Infrastructure*", c'est un portail d'information construit par l'Université Qinghua à Pékin et la Qinghua Dongfang Internet Technology Cie. Ltd. avec un budget initial de 34,5 millions d'euros. Le projet a tout d'abord commencé en juin 1996 avec la publication d'un CD-Rom recensant 2000 journaux académiques de Chine[6]. L'étape suivante, la mise en ligne de la base *CAJ (China Academic Journals Full-text Database)* en 1999 qui comporte alors 3600 journaux s'est déclinée par la suite en de nombreux produits: journaux académiques, thèses, articles, annuaires, etc.

En 2003, elle prend sa forme actuelle de portail regroupant plusieurs bases de données de ressources académiques chinoises avec des fonctionnalités de recherches assez développées : recherches croisées sur plusieurs bases, moteur intelligent, corrélation de connaissances, etc.[7]

Avec actuellement quelque 17 millions d'articles, elle est utilisée en Chine par plus de 5500 institutions: universités, bibliothèques, hôpitaux, organismes gouvernementaux, etc. et ambitionne de devenir le plus gros portail de ressources académiques du monde.

Le développement de cette base démontre bien l'évolution depuis une douzaine d'années de l'Internet et des services associés en Chine. De même, la numérisation des documents prend une ampleur considérable entre 1998 et 2002 et de nombreuses bibliothèques numériques voient le jour. Ainsi, en 1997, le Ministère de la Culture soutient le premier projet d'une bibliothèque virtuelle à Shanghai (*China Pilot Digital Library Project – CPDLP*)[8] à hauteur d'un montant de 110 millions d'euros[9]. En 1998, c'est au tour du Ministère de l'Éducation de porter le projet d'une bibliothèque académique nationale basé sur les systèmes d'information (*China Academic Library & Information System – CALIS*) et du Ministère des Sciences et Technologies de s'atteler à la construction d'un projet de bibliothèque nationale des sciences et techniques. Liu Wei[10], directeur de la bibliothèque nationale de Chine à Shanghai déclare qu'en 1999, 30 millions de pages de documents avaient déjà été numérisées. Le projet *CNKI* s'inscrit naturellement dans la continuité de ces travaux[11,12].

La consultation des notices bibliographiques est gratuite, seul le téléchargement de texte intégral devient payant. Le paiement se fait en ligne via diverses possibilités : une carte de membre, rechargeable à la banque, le téléphone portable, des cartes de crédits. C'est un premier inconvénient pour le chercheur occidental qui ne possède pas de compte en Chine. Il faudra patienter encore un peu pour les paiements en ligne via des sources de paiements internationaux.

Cela dit, les bases sont "propres" au niveau de la structuration de l'information, ce qui rend possible l'analyse par des outils de traitement infomé-

Fig.4 : Page de résultats d'une requête sur CNKI.

triques. Nous entendons par là que les données sont structurées de manière rigoureuse et identique pour chaque article. Les champs de métadonnées sont suffisamment organisés et renseignés pour pouvoir faire l'objet d'un traitement automatisé comme on peut le constater sur la figure 4.

Page d'accueil et ses fonctionnalités

Fig. 5 : Indication des différentes zones de la page d'accueil.

Nous allons expliquer cette page selon les différents blocs qui la constituent (voir figure 5).

Zone 1

Ce champ permet de s'enregistrer avec un nom d'utilisateur et un mot de passe (inscription gratuite) et éventuellement de recharger ses crédits pour l'achat (le téléchargement) de documents complets.

Zone 2

Plusieurs onglets sont ici proposés:

- Présentation de *CNKI*
- Plate-forme de publication numérique (page d'accueil)
- Avis de droit d'auteur
- Analyses statistiques concernant les diverses bases de données couvertes
- Téléchargement des logiciels Acrobat Reader et CAJViewer
- Achats de crédits pour le téléchargement de documents complets
- Contacts, coordonnées
- Support Center
- Aide
- Anglais

Ce dernier onglet ne propose pas l'intégralité du site en version anglaise mais une présentation plus sommaire que celle proposée en chinois. Certaines fonctionnalités seront proposées en anglais, nous le préciserons à chaque fois dans cette présentation du portail.

Zone 3

Ce champ est la contraction d'une page d'accueil simplifiée qui permet d'accéder directement au moteur de recherche. Tant cette page d'accueil que la page de résultats ressemblent au niveau ergonomique à Google. Lancer une requête dans chacun de ces deux espaces correspond à la même action.

Différentes fonctionnalités du moteur de recherche

Fig. 6: Page d'accueil correspondante à la zone 3.

Sur la figure 6, on peut observer 3 fonctionnalités importantes :

1. Cette fonctionnalité permet d'obtenir des résultats en chinois et en anglais pour une requête lancée en une seule langue.
2. Cette fonctionnalité permet d'obtenir des fonctions de recherche avancées. Peuvent être précisés des requêtes sur les champs :

- mot du titre
- mot clef
- date (entre xxxx et xxxx année)
- auteur
- nom de la publication

Fig. 7 : Précision des requêtes.

3. Champ de requêtage

La fonction 2 mène sur la page illustrée sur la figure 7.

Exemple de lancement d'une requête en anglais

Le lancement d'une requête en anglais avec les mots clefs "hybrid wheat" donne 2579 réponses

Fig. 8 : Extrait des réponses obtenues en anglais.

(voir figure 8). La même requête lancée en chinois donne 54061 réponses (voir figure 9).

Plusieurs remarques peuvent d'emblées être établies:

- La première est le différentiel de réponses entre les deux langues; 2 579 articles proposés en anglais pour 54 061 en chinois. La différence est énorme. D'elle, se décline le postulat de ce travail selon lequel il devrait être possible de trouver dans la littérature chinoise des informations n'ayant pas été publiées ou traduites en anglais. Ces informations ne sont donc pas accessibles à un public international[13].
- La deuxième est la constatation que la recherche s'effectue directement dans le corps du texte de l'article. Des options de recherche complémentaires sont possibles, comme l'exigence que le mot clef se trouve dans le titre de l'article par exemple (voir figure 10).

D'autres champs peuvent être sollicités pour la recherche :

- Résumé
- Auteur
- Mot clef
- Titre
- Corps du texte

Présentation des résultats

Sur le côté gauche de l'écran s'ouvre une colonne qui spécifie (voir figure 11).

- la répartition par source des références proposées : Journal académique, conférence, mémoire de master, thèse, publication bilingue chinois / anglais, etc.
- les autres mots-clefs les plus fréquemment cités dans les mêmes documents
- les années de publication

Fig. 9 : Extrait des réponses obtenues en chinois.

Cette forme de présentation des résultats, si elle est ergonomiquement intéressante pour le lecteur (elle se calque sur le modèle Google du point de vue de la présentation) est par contre inutilisable du point de vue de l'exploitation des résultats. Comment trouver les informations intéressantes parmi les quelques 54 000 documents proposés ? Cette interface semble avoir été faite pour satisfaire un utilisateur non-professionnel de la recherche documentaire. Nous verrons plus loin qu'une autre interface, plus structurée et donc utilisable dans une optique de traitement automatique est accessible.

Fig. 10 : Précision de la requête – Recherche uniquement sur le titre.

Fig. 11 : Présentation des résultats.

Zone 4

Cet ensemble contient des liens menant à des thèmes divers, parmi lesquels :

- Ouvrages de référence en ligne
- Navigation sur site universitaire
- Forum académique
- Annuaire chinois de plates-formes d'édition en ligne
- Statistiques sur les bases proposées
- Forum sur des classiques de la littérature chinoise
- Méthodologies (recherche d'information, rédaction de thèse, demande de financement, etc.)
- Réseaux d'anciens élèves et autres réseaux sociaux
- L'innovation dans les entreprises chinoises
- Logiciels de *CNKI*
- Dictionnaire chinois de mots techniques et scientifiques et dictionnaires traductifs chinois/anglais.
- Etc.

Zone 5

Les domaines dans lesquels vont être classés les documents référencés dans le portail sont répertoriés ici. On peut avoir en outre une présentation plus détaillée des différents domaines.

Les domaines sont répartis en neuf grandes classes (voir figure 12) :

- sciences et technologies
- agriculture
- médecine et hygiène
- littérature
- histoire et philosophie
- économie
- sciences politiques et droit
- sciences de l'éducation
- électronique et sciences de l'information

On peut donc, dès la page d'accueil, sélectionner les domaines sur lesquels l'on souhaite focaliser la recherche. Les documents sont ensuite classés dans 6 bases de données qui se déclinent comme présenté dans le tableau 2.

Nous proposons une description focalisée sur la base *CAJ* car c'est d'une part la plus importante en nombre et d'autre part c'est elle qui référence les publications scientifiques les plus intéressantes dans tous les domaines. La *CJP* comportant surtout des archives, ce ne sont donc pas les documents de la science actuelle, la *CDMD* et la *CCND* pourraient se révéler utiles dans une démarche de recherche d'information déjà ciblée (savoir quelle thèse on cherche par exemple parce qu'on aurait déjà lu des articles d'un auteur), la *CPCD* est également intéressante mais là encore, elle est plus pratique dans la mesure où l'on cible une conférence précise d'une part, d'autre part, le laps de temps entre le déroulement de la conférence et la publication des articles est plus ou moins longue selon les conférence. La base *CYFD* recense les livres annuels par domaine d'activités élaborés par les services chinois de statistiques.
La base *CAJ* a de plus l'intérêt de comporter une approche en langue anglaise que n'ont pas les autres.

Nous avons effectué une recherche simple (voir figure 12) en chinois sur le blé hybride dans le

Fig. 12 : Présentation détaillée des domaines.

champ "résumé", en limitant la recherche aux documents publiés entre 2000 et 2008 et en choisissant le domaine "agriculture". Cette recherche donne 1514 articles (voir figure 13). Les mots clefs ressortent en rouge. Le tableau de

Tab. 2 : Détail des contenus des bases de données pouvant être sélectionnées (arrêté au 31 décembre 2007)

Nom de la base		Nature du document	Nombre de titres	Années couvertes	Nombre d'articles
CAJ	China Academic Journals Full-text Database	Publications scientifiques	7,500	Depuis 1994	15,908,460
CJP	Century Journals Project	Publications scientifiques	3,964	Depuis 1887	2,965,927
CDMD	China Doctor/Master Dissertations Full-text Database	Thèses	400	Depuis 1999	322,458
CCND	China Core Newspapers Database	Presse	1000	Depuis 2000	5,845,314
CPCD	China Proceedings of Conference Full-text Database	Conférences	1,018	Depuis 1999	464,127
CYFD	China Yearbooks Full-text Database	Annuaires	1,000	Depuis 1912	2,934,233

résultats affiche les champs "Titre", "Auteur" et "Journal".

En cliquant sur le titre d'un document proposé (fig. 14), une notice plus détaillée est donnée comportant la traduction en anglais de plusieurs

très divers et on imagine mal que la même personne puisse être à la fois expert en génomique, en affaires militaires, en médecine, en agronomie et en géographie (thèmes des journaux proposés dans les premières réponses). La recherche selon le critère "auteur" devra donc être affinée par

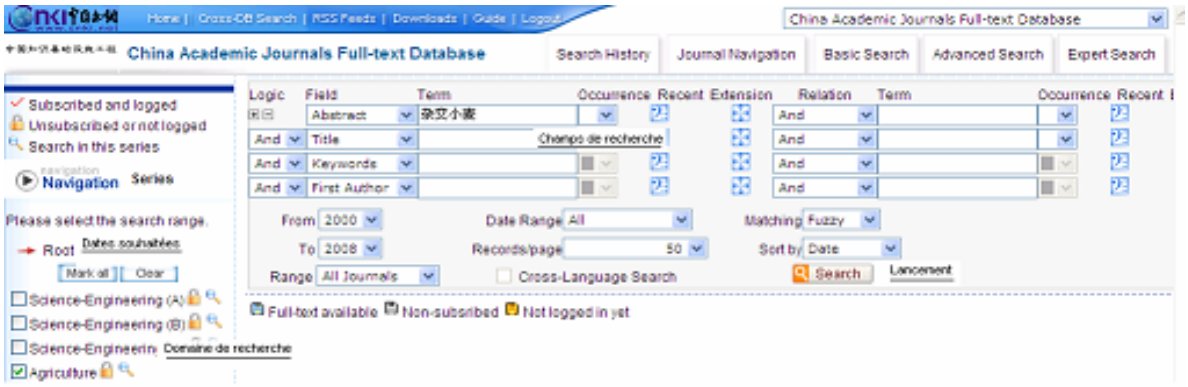

Fig. 13 : Mise en place des éléments de la requête : mots clefs « 杂交小麦 » (blé hybride) devant figurer dans le résumé de l'article dans le domaine agriculture sur la période de 2000 à 2008.

champs : "titre", "auteur", "nom du journal", "mots clef" et "résumé" (fig. 15). Parfois le nom de l'institution est également traduit. Tous les éléments en bleu sont interactifs et constituent des liens vers d'autres pages.

Par exemple, cliquer sur le premier auteur du précédent article (Chen Jie) ouvrira une fenêtre pop-up proposant 4 bases de données où ont été référencés des articles avec pour auteur Chen Jie (voir figures 15 et 16).

Il faudra faire attention à la récurrence des noms chinois ; on parle en Chine des 100 noms de famille (expression consacrée). En réalité, il y en a un peu plus mais c'est quand même très peu au regard de la population). On constate que les 390 articles proposés le sont dans des domaines

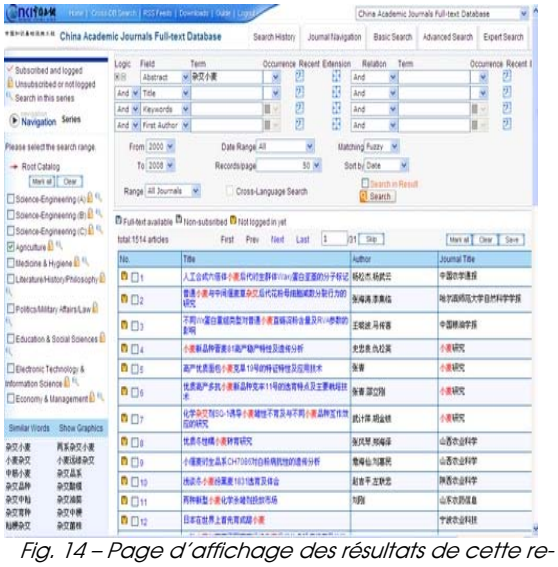

Fig. 14 – Page d'affichage des résultats de cette requête.

croisement avec des mots clefs sur le titre ou l'abstract par exemple.

Fig. 15 : Extraction d'un document.

Pour visualiser sur une seule page les notices de plusieurs articles on peut sélectionner 50 articles (fonction "mark all") et les sauver sur une nouvelle page (fonction "save") (voir figure 18). On peut également utiliser la fonction "customize pour éditer des références formatées (voir figure 19).

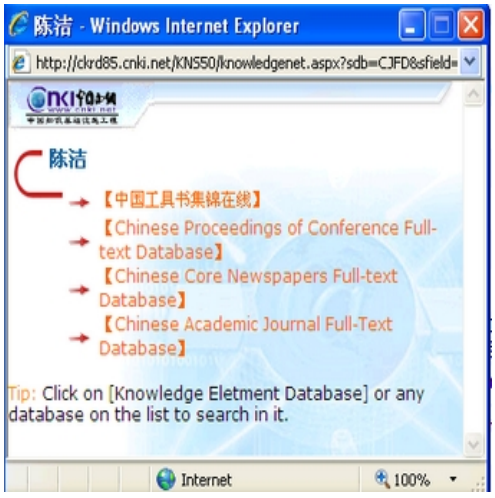

Fig. 16 : Liste des autres bases de données dans lesquelles sont référencés d'autres articles de cet auteur.

Nous obtenons donc maintenant un corpus de résultats qui se présente de manière structuré et homogène contrairement aux résultats proposés par l'interface de la figure 6.

Ce type de données va pouvoir donner lieu à un traitement automatique sur métadonnées par exemple grâce à des outils classiques de traitement automatique de l'information[14]. Le téléchargement du corpus doit cependant se faire manuellement car aucun outil n'est prévu pour cette fonction. Il est de ce fait long et fastidieux et demande une connexion haut débit fiable.

Fig. 17 : Détail des articles publiés dans une des bases présentées à la fig. 16, le nom de l'auteur visé apparait en rouge.

Coût de la recherche

Sur *CNKI*, la connexion et la recherche sont gratuites, le téléchargement coûte 1 à 2 euros par article. Seul problème : les moyens de paiement sont exclusivement chinois : China Construction Bank, Bank of China, China Merchant Bank, China Bank of Telecommunication, China Unicom, etc. Il est donc nécessaire d'avoir un relais en Chine ne serait-ce que pour acheter une carte de membre du portail *CNKI* que l'on peut créditer de 5 à 800 euros (solution la plus intéressante pour un chercheur basé en dehors de Chine).

Fig. 18 : Edition des titres des articles en groupe, sur une seule page.

Conclusion

L'accès, la diffusion et la préservation de l'information scientifique constituent des enjeux majeurs pour la compétitivité d'un pays ou d'une communauté scientifique.

Grâce au projet *CNKI*, la Chine se dote d'un dispositif très performant de stockage, de recherche et de consultation de sa propre production scientifique. Les chercheurs chinois sont habitués à

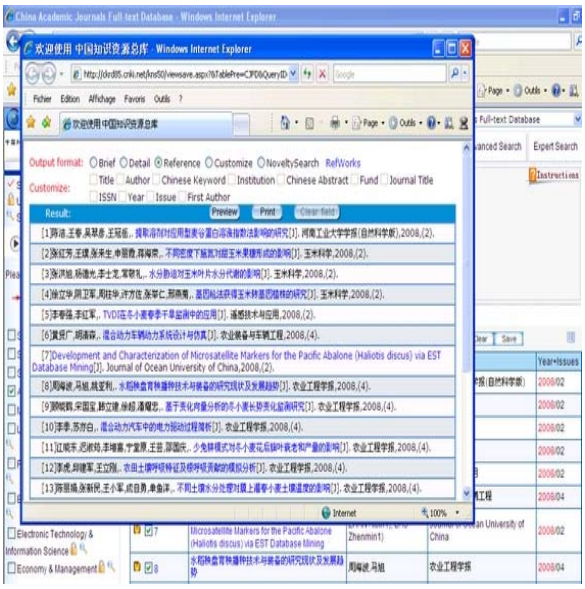

Fig. 19 : Utilisation de la fonction "customize" pour éditer des références formatées.

l'anglais qu'ils maîtrisent au moins dans leur domaine de recherche. Ils ont donc accès à l'ensemble des données mondiales concernant un domaine de recherche (Les universités chinoises ont toutes leurs propres abonnements aux grandes bases internationales (*PubMed, SCI, Biosis*, etc.) et proposent des accès gratuits pour leurs chercheurs).

Comme on le constate depuis 2007, les chinois sont à la pointe de la recherche dans bien des secteurs de la recherche moderne et si le feu vert leur a été donné pour se montrer prolixes sur la scène internationale, la publication dans leur langue maternelle reste plus naturelle et spontanée. Cette dernière est donc largement plus abondante que les publications d'auteurs chinois en anglais. La distinction entre les publications chinoises en anglais et celles en chinois peuvent être motivées par différents aspects :

- Une récompense pour un chercheur : recevoir l'autorisation par sa hiérarchie de publier en anglais, donc de se donner une visibilité sur la scène internationale accroît la fierté personnelle ainsi que la légitimité scientifique et/ou hiérarchique par rapport aux autres membres du laboratoire ou du domaine de recherche au niveau national. (Un laboratoire de recherche universitaire chinois peut compter jusqu'à une cinquantaine de chercheurs). Participer à un colloque international se tenant à l'étranger, c'est se faire payer un voyage "en dehors du pays", ce qui est une haute gratification. (Un chercheur chinois (universitaire) a en moyenne un salaire mensuel de 200 à 500 euros).
- Une stimulation pour les chercheurs (seuls sont traduits en anglais les articles d'un très bon niveau)
- Une protection : c'est ce qui semble avoir été la stratégie du gouvernement chinois avant 2007. Des laboratoires entiers menaient ardemment des recherches dans des secteurs précis depuis une quinzaine d'années mais n'avaient jamais publié à l'international. En 2007, le feu vert leur a été donné et le nombre de publications scientifiques chinoises de haut niveau dans les revues internationales explose[15]. Les chercheurs occidentaux ne s'y attendaient pas et sont étonnés de constater que les chercheurs chinois sont à la pointe de la recherche.
  Ne pas nourrir sa veille des articles scientifiques chinois, c'est donc se couper d'une partie aujourd'hui importante et demain incontournable de l'évolution de la recherche scientifique mondiale.
  Compte tenu du volume, il est bien évident que l'on ne pourra pas tout lire, même sur un sujet précis, surtout avec les difficultés de traduction. Il devient donc nécessaire de proposer une méthode de traitement de cette masse d'articles.
- Par une interrogation ciblée, repérer les articles clefs et les faire traduire avant de les soumettre au jugement d'un expert pour évaluer la qualité ?
- À partir de la recherche, cibler des personnes et des institutions clef et ensuite nouer des contacts ?
- Embaucher dans les laboratoires de pointe occidentaux des personnes ayant un "background scientifique" et parlant parfaitement le chinois (ce qui paraît utopique).
  Une réflexion devrait être menée dans ce domaine, car l'information reste en quelque sorte la clef de l'innovation. Le travail de recherche que nous menons s'atèle à cette dimension de la recherche académique en information scientifique et technique avec notamment des adaptations des outils de traitement automatique de l'information déjà produits par la recherche française (Tétralogie, Matheo, MEVA) qui devront prendre en compte les caractères chinois en se basant sur le codage informatique de ceux-ci.
  En même temps, cela pose le problème de la seule utilisation de sources d'information anglo-saxonnes pour développer une recherche performante. Mais alors, le coût de l'exploitation des bases de données chinoises pose le problème crucial de l'allocation de ressources financières entre autres au niveau de la maîtrise de la langue chinoise. Une suggestion semble possible : pour les domaines de développement stratégiques, entre autres pour les institutions publiques, il serait inté-

ressant de développer un ou des centres de recherche en information spécialisés sur la Chine.

Nadège GUÉNEC
Université Paris-Est Marne la Vallée
Laboratoire S3IS,
Cité Descartes, Champs sur Marne
77454 Marne la Vallée Cedex 2
France
nadegeguenec@gmail.com

Henri DOU
École Supérieure de Commerce et de Management (ESCEM)
1, rue Léo Delibes, BP 0535
37205 Tours Cedex 3
France
douhenri@yahoo.fr

Novembre 2008

Notes

1 BOUTIN, Éric et PEI Liu. Les bases de données en Chine, *Analystes et Décideurs*, n°3, Septembre 2006.

2 CNKI : <http://www.cnki.net/> (consulté le 11 octobre 2008).

3 China Internet Network Information Center. *21th Statistical Survey Report on the Internet Development in China* January 2008.

4 OECD Science, Technologie et Industrie ; perspectives de l'OCDE 2006, 4 Dec. 2006 Pages: 250.

5 BUTLER Declan. China: The great contender. *Nature*, 454, 23 July 2008.

6 LAI, Maosheng ; YANG, Xiudan. Ontology based metadata schema for chinese digital libraries, *5th agricultural ontology service workshop*, FAO, Beijing, April 2004. <ftp://ftp.fao.org/gi/gil/gilws/aims/publications/workshops/AOS_5/ppt/3-1.pdf> (consulté le 11 octobre 2008).

7 LIU, Jia. Metadata development in China, research and practice. *D-Lib Magazine*, 2004, Vol. 10, n°12. <http://www.dlib.org/dlib/december04/liu/12liu.html> (consulté le 11 octobre 2008).

8 WU, Jianzhong. A dynamic gateway to information: electronic services at Shanghai library. In *World library and information congress: 69th IFLA General conference and council*, Berlin, August 2003 (National libraries workshop). <http://www.ifla.org/IV/ifla69/papers/029e-Wu.pdf> (consulté le 11 octobre 2008).

9 Research Information Network. *Activities, costs and funding flows in the scholarly communications system in the UK*, Mai 2008. < http://www.rin.ac.uk/files/Income-&-cost-flows-report-(full-version)-final.pdf > (consulté le 11 octobre 2008).

10 MA, Jiaji. *Bibliothèque et Document Numérique en Chine*, note de synthèse pour le RTP-DOC, CNRS, Juin 2005. <http://rtp-doc.enssib.fr/IMG/pdf/BN_et_DN_en_Chine.pdf> (consulté le 11 octobre 2008).

11 STURGEON, Roy L. Preserving the Past, Preparing for the Future: Modern Chinese Libraries and Librarianship, 1898–2000s. *World libraries*, 2004, Vol. 14, n°1. < http://www.worlib.org/vol14no1/sturg_v14n1.shtml> (consulté le 11 octobre 2008)

12 The Wellcome Trust. *Economic analysis of scientific research publishing.* A commissioned report. January 2003. <http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtd003182.pdf > (consulté le 11 octobre 2008)

13 ZHENG, Shenli. *L'incidence économique de la protection des bases de données en Chine*, OMPI, Comité permanent du droit d'auteur et des droits connexes, septième session, Genève, 13-17 mai 2002. <http://64.233.183.104/search?q=cache:ItiPXxRUsHMJ:www.wipo.int/edocs/mdocs/copyright/fr/sccr_7/sccr_7_6.doc+L%E2%80%99incidence+%C3%A9conomique+de+la+protection+des+bases+de+donn%C3%A9es+en+Chine&hl=fr&ct=clnk&cd=2&gl=be> (consulté le 11 octobre 2008).

14 GUÉNEC Nadège, LOUBIER Eloïse, GHALAMALLAH Ilhème, DOUSSET Bernard. Management and analysis of Chinese database extracted knowledge, *Flexible Query Answering (FQAS 2008)*, Londres, 22 Septembre 2008.

15 Un chercheur chinois dans le domaine de la génomique nous a ainsi confié en mars 2007 qu'il dirigeait des travaux depuis une quinzaine d'années mais qu'il n'avait reçu l'autorisation de publier ses travaux que récemment. Il était auparavant tenu au secret. Nous ne pouvons dévoiler de plus amples précisions sur cette personne eu égard à la confidentialité à laquelle nous sommes tenus.