BASES
POUR UNE RECHERCHE INTELLIGENTE D'INFORMATION
Une publication de FLA Consultants
www.fla-consultants.com

N° 246 - Février 2008

BASES PUBLICATIONS
Commission paritaire 1110 I 87287 • ISSN 0765-1325
SARL de presse au capital de 10 000 €
RC Paris B335325007
Siège social : 27, rue de la Vistule - 75013 Paris
Tél.: 01 45 82 75 75 • Fax : 01 45 82 46 04
contact@bases-publications.com
Directeur de la Publication : François LIBMANN

La parole est d'or... Comment la retrouver ?

François Libmann

Retrouver le discours – sous quelque forme que ce soit – d'une personne, est une problématique relativement fréquente. Les motivations pour cela peuvent être très variées : connaître précisément sa position sur un sujet ou une série de sujets, se faire une idée de sa personnalité à travers ses déclarations, retrouver une citation exacte, voire même chercher des déclarations contradictoires...

La presse et les retranscriptions de radio ou télévisions (surtout aux Etats-Unis) sont les meilleures sources pour retrouver les interviews et déclarations, d'autant que le nombre de titres de presse en ligne ne cesse de croître.

Comme on le verra, outre chercher avec le terme *interview* et/ou le nom de la personne dans le titre, il existe de multiples stratégies – souvent complexes et astucieuses, – permettant de localiser une grande partie des interviews et déclarations d'une personne.

Il faut néanmoins savoir que, si cette personne est connue ou très connue – donc souvent ou très souvent citée –, aucune démarche n'est fiable à 100 %.

Pour une personne peu connue, viser la simplicité

Si la personne est peu connue et si l'on veut être à peu près sûr de ne rien rater, on effectuera tout simplement la recherche sur son nom puis, d'après le titre et/ou un extrait de l'article, on pourra décider s'il s'agit bien d'une interview ou d'une déclaration.

On saisira pour cela le nom et le prénom, séparés d'un voire de deux mots dans les deux sens (opérateur de proximité et non d'adjacence). Dans certains cas, il faudra penser au titre de la personne et chercher l'expression *Titre Nom,* par exemple pour un avocat ou un Président.

Pour une personne connue, de multiples possibilités

Dans le cas d'une personne connue et abondamment citée, pas seulement pour ses déclarations ou à l'occasion d'une interview, de nombreuses stratégies sont possibles.

La parole est d'or... *suite*

De nombeux serveurs/agrégateurs peuvent être utilisés pour ce type de recherche, en l'occurrence ceux qui ont une quantité suffisante de titres de presse en texte intégral et/ou une indexation performante.

Si certaines stratégies peuvent aisément être utilisées sur la totalité des serveurs/agrégateurs – à la traduction près dans une autre langue –, d'autres stratégies font appel à un opérateur spécifique ou à une indexation propre à une banque de données ou à un serveur. Nous présenterons les unes comme les autres.

Sophistication chez EDD

Commençons par EDD, qui propose le texte intégral de plusieurs centaines de titres, essentiellement de presse française.

Cet agrégateur n'offre qu'une indexation un peu rustique par domaine, mais dispose de l'opérateur original "*quorum*" (voir Bases n°216, mai 2005), qui permet de sélectionner les articles contenant au moins *n* termes d'une liste.

Charles Patou, responsable de l'assistance de EDD et son équipe, nous ont suggéré trois stratégies astucieuses qui prennent en compte le registre de langue et les "routines" d'expressions journalistiques.

La première consiste à utiliser des pronoms personnels de la première personne du singulier, en tirant parti de l'opérateur *quorum*, couplé avec une demande de double occurrence d'au moins l'un d'eux.

La stratégie s'écrit *(je: me: moi: ma: mon: mien*)q>=2 et (je ou me ou moi ou ma ou mon ou mien*)>=2*.
Cela permet de capter des articles se rapportant à des déclarations au style direct telles que "*Je pense que ma position à la tête de la mairie découle de mon travail de ces dernières années*".

Les lecteurs attentifs constateront que si les conditions du quorum sont réunies dans cette seule phrase, aucun des termes n'y est présent deux fois. On rappellera simplement que cette double occurrence est demandée dans l'ensemble du document et non sur une seule phrase.

La deuxième astuce consiste à utiliser des codes du langage de la presse. On écrira alors *"dans un communiqué de presse"* ou *"interview de"* ou *"a déclaré"* ou *"a confirmé"* ou *"a réagi"* ou *"a estimé"* ou *"a répondu"* ou *"propos recueillis par"*. Pour cette dernière expression néanmoins, il est utile de la rechercher dans le texte, mais aussi dans le champ *auteur*, en cochant la case correspondante.

On obtient en effet, sur une période d'un mois, 1365 réponses dans le texte et 1790 dans le champ auteur, dont 1464 – soit 82% – uniquement dans ce dernier champ.

La troisième technique consiste à utiliser des pronoms personnels en chapô-titre, pour capter les interviews dont la titraille reprend une déclaration de la personne interrogée. On écrit alors *(je ou nous ou ma ou mon).ti,ch* pour retrouver par exemple *"Mon goût va aux écrivains artistes"*, qui est une conversation avec Jacques Chessex dans *Le Monde*.

Pour bien retrouver ce qui se rapporte à une personnalité donnée, un bon compromis consiste à chercher son nom en titre et chapô (*juppe.ti,ch*) et à combiner avec un ou plusieurs des résultats des stratégies citées plus haut.

On pourra aussi rechercher certaines expressions de la deuxième stratégie à proximité du nom de la personnalité.

Chez Europresse : une fonctionnalité originale

L'approche de Marie L'Haridon, documentaliste au service client d'Europresse, est différente mais tout aussi intéressante. Elle suggère ainsi trois approches complémentaires.

La première consiste à utiliser les clefs de recherche pour retrouver les interviews, quand il existe dans un titre de presse une rubrique s'appelant ainsi.

Si l'on écrit par exemple *CHR_COL=interview*, on retrouvera les documents de cette rubrique dans le quotidien économique *Les Echos* et si l'on écrit *SECT=Interview*, on trouvera des interviews issues de publications telles que *La Vie Française, Le Journal des Finances* ou *Environnement et Stratégies*.

On prendra garde au fait que ces stratégies ne fonctionnent que dans la recherche avancée et pas dans la recherche simple de l'ancienne interface.

En ce qui concerne la nouvelle interface, l'utilisation de ces stratégies est réservée aux abonnés. Ceux qui utilisent la nouvelle interface avec paiement par carte bancaire n'y ont pas accès, ce qui est bien dommage.

La deuxième approche consiste à utiliser les termes et expressions couramment utilisés lorsqu'une interview est reportée dans un article.

On écrira par exemple *propos $2 (recueillis ou reportés) – $n* est l'opérateur de proximité avec l'ordre des mots imposés.

La troisième approche, plus originale, utilise une fonctionnalité propre à la nouvelle interface d'Europresse. Cette dernière permet, en écrivant un + à la fin d'un verbe à l'infinitif, d'étendre la recherche à toutes les formes déclinables de ce verbe.
On pourra donc tenter de retrouver des petites phrases ou citations de personnalités, l'idée étant de rechercher des verbes ou des propositions associés à des propos reportés.

On écrira *(selon ou d $1 apres ou evaluer+ ou analyser+ ou annoncer+ ou prevoir+ ou confirmer+ ou juger+ ou parler+ ou resumer+ ou souligner+ ou constater+ ou indiquer+) %5 Nom de la personnalité*, sachant que % est l'opérateur de proximité, l'ordre des mots étant indifférent.

On pourra compléter ces propositions par des expressions comme *"entretien avec", "propos rapportés", "questions à", "interview"* – ces deux dernières expressions en particulier dans le titre – ou des verbes comme *précise, confie, explique, relativise, martèle, prévient, estime, raconte, met en garde*...

Factiva : une indexation spécifique

Si l'on s'intéresse maintenant à Factiva, la première démarche, la plus simple, consiste à utiliser le *"Factiva Intelligent Indexing"*, c'est à dire l'indexation disponible dans chacun des documents présents sur cet agrégateur.

Deux codes sont pertinents dans ce type de démarche.

NITV est attribué aux articles basés en grande partie sur l'interview d'une ou de plusieurs personnes et à ceux qui se présentent sous la forme de "questions réponses".

NTRA est attribué à une reproduction écrite de ce qui a été dit lors d'une conférence de presse ou lors d'un passage à la radio et/ou à la télévision.
Il peut s'agir de passages d'un discours ou d'une interview.
Ce code n'est pas attribué lorsqu'il s'agit seulement de brèves citations dans un article.

Ces codes ont été appliqués à partir du 1er mai 2000.
Pour utiliser les deux codes ensemble, il faut écrire *ns=NITV OR ns=NTRA*.

L'un des intérêts de ces codes est qu'ils s'appliquent à des documents dans toutes les langues, français et anglais bien sûr, mais aussi allemand, italien, russe, japonais, chinois...

L'inconvénient en revanche est que l'on ne peut pas mettre d'opérateur de proximité entre le nom de la personne interviewée et le code.

Pour augmenter les chances de retrouver des interviews ou déclarations de personnes qui nous intéressent, on pourra mettre leur nom dans le titre et/ou le paragraphe principal (HLP), ou dans l'ensemble des titres/sous-titres/surtitres (HL) ou encore dans les titres (HD), en écrivant par exemple *HD=nom* ; mais il faut reconnaître que le résultat n'est pas parfait.

Un test en anglais sur un an donne 67 966 documents indexés avec NITV ; parmi ceux-ci, 36 585 – soit plus de la moitié – ont *interview* dans le titre (*HD=interview*).

A l'inverse, 54 870 documents comportent le mot *interview* dans le titre ; mais 18 285 d'entre eux – soit un tiers – ne sont pas indexés avec NITV, un nombre non négligeable étant pourtant de réelles interviews.

En complément de l'utilisation de ces codes, on pourra utilement s'inspirer de certaines méthodologies présentées sur les agrégateurs précédents.

En anglais, on pourra utiliser des termes ou expressions comme *said, says, predicts, predicted, according to, talks, talked, warns, warned, explains, explained*... à proximité du nom de la personnalité, avec l'opérateur *adj5* ou *near5* selon que l'on exige l'adjacence dans cet ordre ou la simple proximité.

On peut aussi chercher dans le titre *questions, a word with, Q&A, interview*... On sera, par ailleurs, attentif au fait qu'en anglais, il est plus fréquent qu'en français d'utiliser le nom de famille sans le prénom.

LexisNexis : privilégier la bibliothèque News

En ce qui concerne LexisNexis dans la bibliothèque *News* en anglais, nous avons effectué des tests sur le fichier *Current News*, qui couvre les deux dernières années.

Comme dans Factiva, l'utilisation de l'indexation donne des résultats mitigés.
Par exemple, si l'on cherche *Fillon* à deux mots maximum d'une série de termes de la liste ci-dessus, on trouve 1092 documents dont 161 seulement sont indexés avec *Interviews* (on écrit *TERMS(INTERVIEWS)*).
En examinant les 935 autres, on découvre plusieurs documents qui sont au moins pour une part des interviews.

Nous avons procédé à un autre test : on trouve dans ce fichier 75 535 documents ayant le mot *interview* dans le titre, mais seulement 48 946 ont à la fois le mot *interview* dans le titre et l'indexation ; ce nombre est porté à 49 190 si l'on cherche *interview* dans le titre ou la section et dans l'indexation.
On en trouve donc 24 345 ayant *interview* dans le titre ou la section mais pas dans l'indexation... Pourtant, là encore, de nombreux articles sont réellement des interviews.

Renseignements pris, l'algorithme sélectionnant les documents indexés avec le terme *interviews* est en cours de révision ; dès le mois d'avril, cette indexation devrait laisser moins de silence..

Nous avons pensé restreindre la recherche à certaines bibliothèques comme PEOPLE (les news sans les biographies), MARKETING ou ENTERTAINMENT, LexisNexis étant connu pour avoir d'excellentes et nombreuses sources dans ce dernier domaine.

Pour voir quel intérêt il pouvait y avoir à interroger telle ou telle de ces bibliothèques, nous avons utilisé une stratégie simple et brutale qui, si elle ne donne pas des résultats exhaustifs, fait preuve d'une certaine efficacité.
Cette stratégie consiste à chercher le mot *interview* dans le titre *(Headline(interview))* pour les publications du 11 février 2008. Les résultats sont assez surprenants. On trouve en effet 124 documents dans *News*, 4 (!) dans *People*, 11 dans *Entertainment* et 49 dans *Marketing*.

Dans cet échantillon, les documents extraits des bibliothèques spécialisées se retrouvaient aussi dans *News*.

Renseignement pris, la quasi totalité du contenu des bibliothèques spécialisées est en fait présent dans la bibliothèque *News*. Il est donc préférable d'interroger celle-ci directement, sauf pour d'eventuelles raisons de coûts.

Dialog : une approche globale ou par base

Quant à Dialog, on sait que le serveur propose plus de 500 banques de données offrant aussi bien des informations news et business que scienfitiques et techniques (littérature et brevets).
Nous nous limiterons ici à ce qui concerne les news et business, ce qui représente plusieurs centaines de banques de données.

Dans d'autres bases, en médecine, psychologie ou sociologie par exemple, le terme *interview* se réfère plutôt à des méthodologies de recueil d'information et est donc hors de notre sujet.

Il y a deux façons d'aborder la recherche dans Dialog.
La première consiste à sélectionner quelques banques de données ayant des indexations spécifiques – par exemple le code Document Type (*DT=interview*) – et à se limiter à ces bases.

La deuxième stratégie consiste à utiliser le Dialindex pour sélectionner les bases qui répondent le mieux, parmi celles des catégories news et business. Différentes stratégies sont alors possibles.

Si l'on souhaite utiliser la première méthode, on pourra sélectionner les bases 148 (Gale Trade & Industry DB), 47 (Gale Group Magazine DB), 484 (Periodical Abs Plustext), 483 (News Papers Ab Daily), 15 (ABI Inform), 553 (Wilson Business Ab), 727 (Canadian Newspapers), 141 (Readers Guide) et 648 (TV and Radio Transcripts).

Ces banques de données sont, dans l'ordre, celles qui répondent le mieux à la stratégie *DT=interview et PY=2007*.

Ces bases ont aussi un champ NM=, dans lequel on mettra le nom de la personne. On pourra aussi la chercher dans le titre et, dans le fichier 15, dans le résumé (pour éviter le texte intégral). On signalera enfin que les fichiers 148 et 47 comportent un nombre non négligeable de simples citations, sans résumé ni texte intégral.
C'est le cas d'un tiers environ des articles trouvés avec *DT=interview*.

Pour illustrer le côté parfois aléatoire de l'indexation – surtout s'il s'agit de l'indexation du même article dans deux bases différentes –, on citera l'exemple d'un article du fichier 211 de Gale, dans lequel on trouve la citation seule (titre + indexation) d'un article du *New York Times,* contenant le mot *interview* à la fois dans le titre et dans le champ *Document Type*. Dans le fichier New York Times FullText (n°471), cet article est surtitré *Saturday Profile* et contient le mot *biography* dans le champ *Document Type*. Il existe par ailleurs dans ce même fichier, à la même date, un article surtitré *Saturday Interview* avec *interview* dans le champ DT.

Pour la deuxième méthode, si l'on veut être le plus large possible on écrira, une fois dans le fichier 411, *SF ALLNEWS ALLBUSINESS* (*SF* étant utilisé pour *Select Files*).

C'est une sélection un peu large mais qui permet de n'omettre aucun fichier intéressant. De toutes les façons, les fichiers scientifiques par exemple n'ont aucune raison de donner des résultats dans une stratégie incluant un nom.

On pourra utiliser deux types de stratégies :
• *S(DT=interview OR interview?/DE OR interview/TI)* et le nom de la personne ;
• une stratégie avec *says OR said OR warns*... comme vu plus haut.
Pour limiter le bruit, on peut exiger que le nom soit dans le titre (attention au nom seul, s'il est ambigü, comme *Jobs*), les descripteurs ou le champ NM, vu plus haut.

Sur DataStar, les temps de réponse du CROS – équivalent du Dialindex de Dialog – étant rédhibitoires dès que le nombre de banques de données interrogées est important, on sélectionnera des bases telles que INDY (équivalent de Trade & Industry), MAGS (équivalent de Magazine DB), INFO (ABI/Inform).

On pourra aussi chercher dans des titres de presse dans d'autres langues que l'anglais, en s'inspirant de certaines des stratégies vues plus haut.

Cette problématique est un bon moyen d'illustrer l'apport des stratégies complexes par rapport à des requêtes simples, même si ces dernières donnent malgré tout des résultats.