



**Compte-rendu du séminaire ELP – Sinequa**  
**10 juillet 2003**

**L'intégration d'un moteur de recherche et de navigation  
sémantique dans les bases documentaires Adhoc**

**Intervenants :**

Sinequa :	Laurent Le Foll, DGA Luc Manigot, Directeur Scientifique
ELP :	Jean-Jacques Bec, Directeur Général Thierry Gauthier, directeur technique Bernard Charnomordic, responsable des programmes documentaires



## **Introduction**

Adhoc, outil développé par ELP pour les applications documentaires, est un gestionnaire de contenu qui s'intègre au cœur du système d'information des entreprises. Puissant outil d'organisation de l'information, Adhoc est accessible via un navigateur.

En deux ans, le nombre d'internautes mécontents de la qualité des moteurs de recherche a doublé en raison de la surabondance de l'information électronique. Les utilisateurs perdent de plus en plus de temps à localiser, catégoriser et analyser l'information disponible. La nécessité de trouver des solutions est indispensable pour permettre aux utilisateurs d'identifier et d'utiliser l'information qui leur est nécessaire. Ces solutions doivent être facilement déployables et hautement personnalisables.

Dans le cadre d'un partenariat technologique, Sinequa et ELP ont œuvré pour la mise en place de la nouvelle version du site internet de la COB. Cette collaboration illustre parfaitement l'intégration d'un outil documentaire et d'un moteur de recherche sémantique afin de structurer un site Web et permettre ainsi d'accéder à l'information pertinente.



## Présentation de Sinequa

Partenaire de ELP, Sinequa est une société d'édition de logiciel, spécialisée dans l'accès intelligent à l'information. Bénéficiant de près de 20 ans d'expérience dans le domaine de la linguistique, Sinequa a développé un logiciel de recherche et de navigation sur le web : **Intuition**. Parmi les utilisateurs exploitant le moteur Intuition sur leur site Internet, on peut citer : Le Monde, Ouest France, La Redoute, Diva Press, Allo Ciné, Cdiscount.com., Lyreco. La SNCF, EDF, Thales, le SIPJ, le Ministère de la Défense et bien d'autres l'utilisent sur leur site Intranet.

Le moteur Intuition est disponible dans la plupart des langues européennes ainsi que dans 3 langues asiatiques. Intuition allie l'analyse syntaxique (analyse lexicale et grammaticale), sémantique et documentaire (utilisation des thesaurii et plans de classements).

La dernière version d'Intuition : Intuition version 4, est compatible avec le format XMP d'Adobe afin de permettre aux entreprises d'accéder au « semantic web ». La norme XMP est basée sur des métadonnées XML. Cette intégration permettra une indexation indépendante du format de fichier (image, PDF) et une diffusion de contenu indépendante du support (papier, web, livre électronique). En outre, les bases de connaissance d'Intuition version 4 seront compatibles avec la norme RDF.



## Le site web de la Commission des Opérations de Bourse

L'association entre ELP et Sinequa vient de 2 constatations : le documentaire ne suffit pas toujours à organiser l'information et la recherche sémantique seule ne permet pas de retrouver facilement sans structuration préalable une information pertinente.

Dans le cadre du projet de la COB, Intuition est venu compléter Adhoc Premium, l'outil de gestion et d'organisation d'information d'ELP.

Les missions fondamentales de la COB sont de garantir la transparence financière auprès du public et de rendre accessible l'information relative aux marchés financiers. Dans le souci d'amélioration de ces services, la COB a décidé la refonte de son site Internet et a choisi d'appuyer l'organisation du contenu du site sur une base documentaire Adhoc.

Afin de faciliter l'accès à l'information du public et d'assurer une meilleure transparence de l'information, ELP a proposé à la COB d'ajouter la puissance du moteur de recherche et de navigation sémantique Sinequa à la recherche documentaire Adhoc.

Sur le site de la COB ([www.cob.fr](http://www.cob.fr)), l'information est structurée comme dans une base Adhoc. Les domaines, fichiers, sous-fichiers de la structure de la base, deviennent des sections, rubriques, sous-rubriques du site web. Chaque partie du menu correspond à des données dynamiques. Ainsi, la rubrique réglementation devient accessible par type de texte : lois, décrets ou règlements et par thème : opérations financières, OPCVM, offres publiques... En parallèle, la réglementation est aussi représentée comme un plan de classement électronique, comme les doctrines de services et les publications de la COB. Le site internet est géré de manière dynamique et peut être modifié en temps réel selon les modifications et informations saisies dans la base.

L'exploitation optimale de l'organisation du site se fait grâce à Adhoc et facilite la navigation des internautes.

Au niveau de la recherche, Adhoc permet de pondérer les recherches effectuées avec le moteur de recherche Intuition.



Grâce au thésaurus boursier de la COB qui comprend plus de 4 000 termes, le langage boursier et la hiérarchie des termes existants sous Adhoc sont fournis au moteur de recherche.

On obtient ainsi une reconnaissance des concepts dans un ensemble de résultats et une meilleure exploitation des structures d'organisation des documents.

L'utilisation combinée du documentaire et de la recherche sémantique permet d'accéder beaucoup plus efficacement et intuitivement à l'information pertinente.



## Présentation de la technologie Intuition

La technologie Intuition repose sur une approche interdisciplinaire : la syntaxe, la sémantique, le documentaire, les statistiques et les métadonnées.

- La syntaxe est prise en compte par l'analyse lexicale et grammaticale ainsi que la reconnaissance des entités et l'extraction des relations.
- La sémantique s'exprime par l'indépendance des mots de la requête et de la langue. Sinequa utilise un concept unique dans la représentation mathématique du langage humain. Intuition permet la neutralisation des pluriels, l'identification des racines et l'analyse des groupes nominaux.
- Le documentaire est retrouvé dans la réutilisation des ressources existantes et la catégorisation dans un ou plusieurs plans de classement.
- Les statistiques sont utilisées dans l'identification des relations cachées, le text-mining.
- La notion de métadonnées apparaît lors de la recherche et l'indexation en données extratextuelles définies par l'utilisateur, l'interrogation et la mise à jour en SQL ainsi que la compatibilité XMP et le support complet de XML et partiel de XPATH.

Toutes ces intégrations apparaissent dans les fonctionnalités développées pour les clients de Sinequa. Le questionnement devient simple et convivial et peut s'effectuer en un mot mal orthographié ou par une phrase complexe dans la langue préférée. Les résultats sont affichés par une réponse riche et contextuelle avec une représentation des concepts et une utilisation des ressources documentaires.



Les spécifications techniques d'Intuition sont les suivantes :

- Architecture trois tiers
- Indexation incrémentale et en temps réel
- Possibilité d'avoir jusqu'à 80 utilisateurs par processeur
- Load balancing et réplication dans les cas de haute disponibilité
- Serveur disponible pour Windows 2000/NT/XP, Sun, Linux,
- Client accessible d'un browser,
- Indexation des fichiers MSOffice, PDF et HTML,
- Support natif de XML (XPath) et de XMP,
- Connecteurs vers SGBDR's, API's : C/C++, ASP, ODBC, PHP, Perl, Java, JDBC, Web Services.



## **Conclusion**

Le site internet est dynamique modifiable en temps réel selon les modifications apportées dans la base Adhoc.

La gestion dynamique d'un site internet au moyen d'une base Adhoc facilite toutes les opérations de mise à jour qui deviennent exécutables sans mettre en cause la cohérence globale du site.

Tout l'effort d'organisation et de structuration des données vient préciser les stratégies de navigation et de recherche du moteur sémantique.

Les technologies documentaires Adhoc, alliées à la puissance du moteur sémantique Sinequa permettent d'apporter des gains très spectaculaires dans l'accès à l'information et aux documents pertinents d'un site WEB.