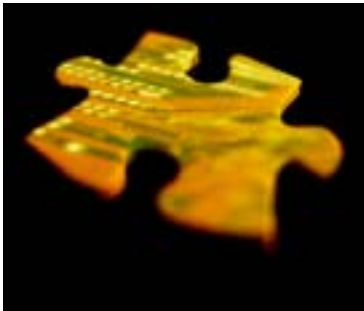
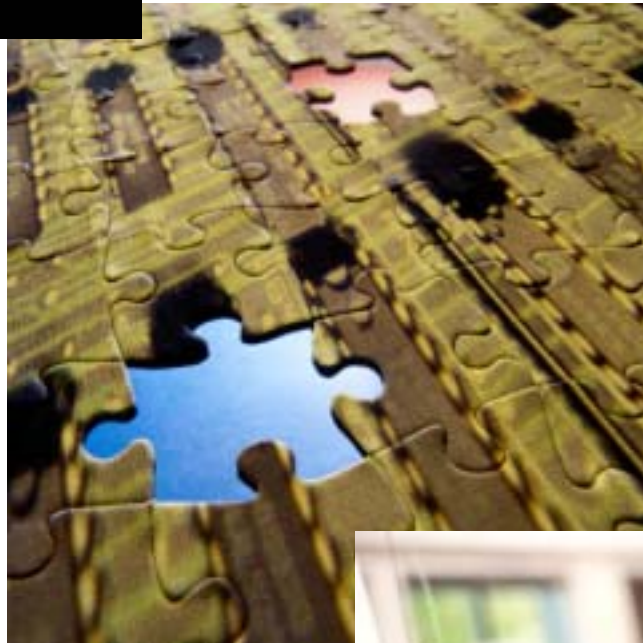




Ingénierie de la connaissance

La recherche d'information multimédia et
multilingue



Point presse du
24 novembre 2004



Ingénierie de la connaissance

La recherche d'information multimédia et multilingue

Des besoins croissants

Beaucoup d'utilisateurs d'internet pensent que le problème de la recherche d'information est résolu parce que l'on peut interroger une grande partie du Web au moyen de moteurs de recherche.

En fait, ces systèmes indexent très vite de gros volumes au détriment de la pertinence de la recherche. Les particuliers les apprécient pour leur rapidité, mais les entreprises ont des besoins plus pointus.

Veille stratégique, commerciale, scientifique et technique, la gestion des connaissances dans l'entreprise, la vidéo à la demande, la protection de la propriété intellectuelle, la formation personnelle et à distance, la localisation de logiciels (adaptation linguistique et culturelle) sont autant d'applications qui nécessitent la mise en œuvre de technologies avancées de recherche d'information.

Les enjeux de l'ingénierie de la connaissance

La mise en œuvre des connaissances est un enjeu stratégique majeur dans tous les domaines de l'activité économique :

- La perte de temps et d'information stratégique peut avoir des conséquences économiques graves car la valorisation d'une entreprise (entre autres sa cotation en bourse) repose de plus en plus sur des actifs immatériels dont son patrimoine de connaissances ;
- La productivité n'est plus seulement relative à l'investissement et la force de travail, mais aussi à la dynamique de mise en œuvre des connaissances (celles qui sont internes à l'entreprise comme celles de la concurrence) ;

- La dimension internationale de certaines entreprises, impose la nécessité de conserver et d'exploiter les connaissances communes entre différents établissements, ayant des cultures différentes, parfois implantés dans des pays différents (donc avec des langues courantes différentes même si l'anglais est souvent utilisé comme langue standard d'échange international) ;

- Pour maintenir leur capacité d'innovation, les entreprises et les organismes de recherche ont besoin de surveiller les évolutions et les avancées dans leurs domaines d'activité. La mondialisation nécessite une surveillance tous azimuts de documents



variés issus des médias (presse, télévision, radio,...) et de sites internet (entreprises, publications scientifiques, sites institutionnels, etc.) ;

- La multiplication des volumes et des supports d'informations, en particulier grâce à internet et aux technologies numériques, pose de manière aiguë le problème de leur recherche et de leur filtrage.

Le CEA List¹, un des experts mondiaux dans le domaine de la maîtrise de l'information

Le Laboratoire d'ingénierie de la connaissance multimédia multilingue (Direction de la recherche technologique, CEA List) maîtrise des technologies dans la recherche et le filtrage d'information multimédia (texte, parole, images fixes et vidéo) multilingue. Ces technologies intéressent les domaines civil et militaire : extraction d'information, résumé automatique de textes, organisation des connaissances à partir de documents de référence, mémoires de traduction, etc.

Pionnier dans le traitement informatique de la langue et de l'image, le CEA a développé depuis les années 70, des outils de recherche d'informations pour produire des connaissances. Comme tous les acteurs de la haute technologie, les équipes du CEA doivent disposer d'outils performants leur permettant de suivre au niveau mondial l'évolution des connaissances scientifiques sur des domaines pointus et de gérer, au sein même de l'organisme, le patrimoine de connaissances qu'ils accumulent.

Certaines des technologies d'ingénierie de la connaissance développées au CEA ont déjà fait l'objet d'une industrialisation et d'une commercialisation à partir des années 1980 et une centaine d'utilisateurs sont équipés des technologies CEA dont la Cour des comptes, la Banque de France, TF1... Ainsi, l'ensemble des arrêts des Cours des comptes nationale et régionales sont accessibles au public par internet via un moteur de recherche basé sur les technologies du CEA.

Un potentiel industriel important : naissance de New Phenix

L'importance des marchés couverts par les enjeux de l'ingénierie de la connaissance conduit à la création de la start-up New Phenix. Cette société exploitera un portefeuille de licences du CEA dans le domaine de l'ingénierie des connaissances et développera des produits de gestion de l'information multimédia multilingue destinés aux entreprises. Elle poursuit une collaboration étroite avec le CEA List de façon à maintenir une offre innovante dans les systèmes génériques pour l'ingénierie de la connaissance.

¹ Le laboratoire d'intégration des systèmes et des technologies (List) est implanté en Ile de France

Des systèmes opérationnels développés au CEA List

- **Gestion de connaissances en huit langues**

Bien que l'anglais soit couramment utilisé pour échanger de l'information entre différentes implantations d'une même entreprise ou même dans le cadre d'une collaboration entre deux entreprises, l'expérience montre que les documents les plus riches en informations technologiques et stratégiques sont dans la langue d'usage des rédacteurs.



Les équipes du CEA List ont développé un logiciel de gestion de la connaissance multilingue fonctionnant dans huit langues différentes : Français, Anglais, Allemand, Espagnol, Italien, Russe, Arabe et Chinois. A partir d'une seule question, il permet de trouver tous les documents pertinents quelle qu'en soit leur langue. Ainsi, une interrogation en français peut permettre d'obtenir un document pertinent écrit en russe ou en arabe. Couplé avec un logiciel de traduction automatique (le logiciel grand public Systran²), le système peut même fournir à l'utilisateur une compréhension minimum (traduction approximative) des documents dans

sa langue usuelle.

Le système peut être mise en oeuvre par une entreprise internationale pour :

- *gérer ses propres documents dont le formalisme et la collecte peuvent être relativement homogènes ;*
- *réaliser une veille stratégique sur la partie publique des connaissances des concurrents, à partir de documents produits et gérés par d'autres opérateurs, dans des formalismes divers.*

- **Veille stratégique, scientifique ou de sécurité**

La compétition internationale impose aux entreprises de disposer d'outils de filtrage d'information, de classification des documents, d'analyse des corrélations entre informations de sources différentes (et dans des langues différentes).

Les technologies développées au CEA List sont mises en œuvre pour des applications dans le domaine de la sécurité publique : lutte contre le terrorisme, contre la grande criminalité, surveillance de la prolifération d'armes de destruction massive. Par exemple, il est possible d'identifier des corrélations entre des documents d'origines différentes (langue et support différents) qui révèlent, lorsqu'ils sont rassemblés, une information stratégique.

² Systran est le numéro un mondial des logiciels de traduction automatique.

- **Traitement de la parole**

Pour traiter les informations sonores, la technologie du CEA List a été couplée à celles du Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur (Limsi) du CNRS et à celles de la société Vecsys qui commercialise les logiciels développés au Limsi.

Il est ainsi possible de générer automatiquement, voire en temps réel, des documents texte à partir de sources sonores.

- **Recherche d'images et de vidéos**

Le marché de la vente de photos et de vidéos s'est considérablement modifié ces dernières années par la vente des données numériques en ligne à un prix forfaitaire. Les éditeurs légendent succinctement leurs photos, en anglais ou dans leur langue courante. Pour faciliter l'exploitation de ces ressources, le CEA List a développé un système d'interrogation interlingue (une seule question dans une langue et toutes les photos trouvées quelle que soit la langue de description) couplé à un second système de recherche par traitement d'images.

Dans un premier temps, l'utilisateur recherche des images à partir d'une requête écrite. Par exemple, à partir du mot « voiture », il trouve des centaines de clichés possédant ce mot (ou un synonyme ou une traduction) dans leur légende. Il peut ensuite retenir un cliché en particulier : par exemple une voiture rouge vue de face. Le système peut alors lui fournir les images similaires.

Cet outil peut être utilisé dans toutes les recherches iconographiques (presse, studio de création graphique, maquettistes, décoration) mais aussi pour la recherche de plagiats d'œuvres d'art.

- **Indexation automatique d'images**

Les producteurs de biens de grande consommation sont intéressés par l'indexation automatique d'images pour le suivi des ventes d'un produit. Lors d'une enquête de consommation, l'enquêteur photographie un rayonnage. Le système référence automatiquement l'emplacement, le nombre et le type d'article présents dans le rayonnage dans une base de donnée.

Un tel outil, basé sur les technologies du CEA List, représente un gain de temps important par rapport au référencement manuel pratiqué actuellement.

Les étapes de la recherche sur des images

Première étape :
recherche sur les
légendes d'images

Requête de l'utilisateur formulée en français : **Couchers de soleil**
Réponse : liste d'images dont la description textuelle correspond quelle que soit la langue :



Coucher de soleil, mer, rochers...



Anocheceer, arboles, sol



Sunset, mountains, ...



Seconde étape :
recherche d'images
semblables à l'image
choisie

Requête : l'utilisateur choisit une image



Réponse : série d'images semblables sur le plan de la couleur



Les outils de la recherche d'information du CEA List

• Recherche textuelle/parole

Pour garantir la plus grande efficacité, il est nécessaire de mettre en place sur les données de type texte ou parole un **traitement linguistique** élaboré comprenant une analyse morphosyntaxique (reconnaître un mot, identifier sa catégorie : nom, adjectif, adverbe). Cette analyse permet une représentation normalisée des concepts et les relations syntactico-sémantiques qui les lient (« compréhension » du sens des mots en fonction de leur utilisation et de leur position syntaxique dans la phrase : par exemple il faut pouvoir distinguer le sens du mot « temps » dans les expressions « prévision du temps » et « temps de calcul »).

Cette analyse permet aussi de reconnaître les entités nommées, de reconnaître et normaliser les dates, ainsi qu'identifier les mesures, en particulier financières. Pour un texte donné, l'outil de recherche mis au point par le CEA List peut :

- o reconnaître que tel mot est un nom de société, de personnes, de lieu, de produits ;
- o normaliser les différentes expressions d'une date comme « 20 octobre », ou « 20/10 », ou « *october the 20th* » et trouver la date précise lorsque le texte ne mentionne qu'« hier » ;
- o distinguer qu'un chiffre indique une longueur, un poids ou une somme en euros par exemple.

Pour atteindre le maximum d'efficacité, cette analyse doit être pratiquée aussi bien sur les documents analysés que sur les requêtes de l'utilisateur.

• Recherche image/vidéo

Les équipes du CEA List développent des systèmes qui permettent de calculer des paramètres de ressemblances d'images. Les traitements portant sur la couleur, la forme et la texture tant au plan local que global de chaque image. Ils permettent d'en donner une représentation qui va servir de description synthétique pour la recherche.

A terme, il sera aussi possible de réaliser automatiquement une reconnaissance d'objets à l'intérieur des images et d'en générer une description textuelle. C'est-à-dire qu'il ne sera plus nécessaire de décrire manuellement par des mots une image car le logiciel le fera automatiquement. Le système fonctionne déjà sur des images simples (quand les objets sont bien séparés, sur un fond uni, comme c'est le cas dans l'exemple ci-contre) et la recherche se poursuit pour étendre son application à tout type d'images.



"Photo de dimension 384 par 307 pixels. Le fond de cette photo est noir. Elle est composée de 3 objets. Cette image a pour thème principal Jouet. L'objet principal est un nounours beige qui se trouve à gauche de la photo. Nous distinguons, au centre de l'image, une balle verte et blanche. Nous notons un cube multicolore à droite."

D'autres applications développées le CEA List

- recherche d'information sur le **Web visible et le Web invisible**³ à l'aide d'un métamoteur⁴ utilisant traitement linguistique et information lexicale ;
- filtrage d'information à la volée sur des **files de presse** ou à partir de recherches sur le Web ;
- **classification** automatique multilingue et classification d'images ;
- alignement de phrases pour les mémoires de **traduction** : une traduction automatique n'est jamais de très bonne qualité. Il faut donc que l'homme révise le document traduit automatiquement. Mais lorsqu'on réalise des traductions successives (notamment les traductions des documentations de produits qui se succèdent) il est possible de conserver la mémoire des anciennes traductions. Un outil informatique aligne les textes et recherche ceux précédemment traduits. Dans le cadre du test comparatif international Arcade réalisé en 1998 et financé par l'Agence universitaire de la Francophonie, la technologie du CEA est apparue comme la plus robuste car ses performances restent constantes quelle que soit la nature des documents à traiter.
- **résumé automatique** de texte mono et multilingue, orientée selon un profil défini par l'utilisateur ou selon la pensée de l'auteur ;
- extraction de terminologie dans des corpus : il s'agit d'extraire, dans un domaine particulier, le **vocabulaire significatif** utilisé, pour améliorer la prise en compte du domaine par les systèmes de traitement linguistique ;
- **Description automatique d'images** en langage naturel.

³ Les pages Web invisibles sont des pages qui n'existent pas mais qui sont créées en dynamique par des logiciels et notamment d'autres moteurs de recherche. On peut ainsi consulter des bases de données en créant des interrogations automatiques.

⁴ Moteur de recherche à mots clés qui interroge les autres moteurs et synthétise les réponses.

Recherche technologique sur les systèmes à logiciel prépondérant



■ Le List développe des technologies innovantes associant le logiciel et le matériel pour la conception et la maîtrise de la complexité des systèmes à haut niveau d'intégration. Laboratoire de recherche technologique, le List met en synergie la R&D industrielle et la recherche fondamentale, dans une dynamique qui va du concept de système jusqu'au démonstrateur pré-industriel.

■ La culture projet, forte et bien établie, des 300 chercheurs, ingénieurs et techniciens du LIST fait de celui-ci le partenaire naturel des industriels dans l'émergence de nouvelles technologies. De longue date, les chercheurs, ingénieurs et techniciens du LIST ont su répondre aux exigences des industriels du nucléaire, en termes de performance, de sûreté, de délai. Ils étudient et développent, aujourd'hui, des solutions adaptées aux besoins des industriels de nombreux secteurs de l'économie, notamment l'automobile, l'aéronautique, la défense, le médical et la communication.

■ Contribuant au transfert de technologies innovantes, les 300 chercheurs, ingénieurs et techniciens du LIST favorisent la création de richesses notamment par l'émergence d'entreprises. Les technologies du List ont fait l'objet de la création de plusieurs start-up, essaimées du CEA, dont Haption (interfaces à retour d'effort, dites « haptiques », Ile-de-France) et Acti-CM (systèmes de mesure 3D, Rhône Alpes).

■ Les équipes du List sont en relation avec de nombreux laboratoires universitaires, des grandes écoles et d'autres organismes de recherche (CNRS, Inria, etc.), au travers de collaborations intégrant des compétences complémentaires. Elles sont aussi reconnues au niveau international (expertise, projets européens, conférences et comités d'experts et de normalisation).

■ Le List fédère les équipes de la Direction de la recherche technologique du CEA basées en Ile-de-France, à Fontenay-aux-Roses (92) et à Saclay (91). Il est porteur, en partenariat avec le département des sciences et technologies de l'information et de la communication du CNRS du projet Num@tec, plate-forme d'innovation pour les technologies numériques, destiné à constituer en région Ile-de-France un pôle de recherche technologique de dimension européenne centré sur les « systèmes à logiciel prépondérant ».

CEA List

18, route du Panorama - BP 6 - 92265 Fontenay-aux-Roses Cedex
Tél. 01 46 54 99 37 - Fax. 01 46 54 99 42



- Laboratoire d'intégration des systèmes et des technologies

Les recherches en ingénierie de la connaissance au CEA List

L'interaction homme – système, au cœur des systèmes intelligents de demain, est déjà omniprésente dans la vie quotidienne des citoyens, dans l'entreprise (nouveaux modes de conception et de production étendus) et dans ses autres activités (énergie, transport, santé, domotique, loisirs...).

Le déploiement des systèmes interactifs et des services passe par la maîtrise de leur complexité croissante. La convivialité et la simplicité d'utilisation d'interfaces de plus en plus intelligentes et transparentes en est la condition essentielle. On parle alors d'intelligence ambiante et d'interaction naturelle et multimodale. Cela explique que l'interactivité, située au centre d'un triptyque homme – information - environnement, soit le pivot de nombreux programmes de recherche dans le monde.

En particulier, les équipes du CEA List mettent en œuvre des technologies innovantes de recherche et de filtrage d'information multimédia multilingue (applications aux domaines civil et militaire), d'extraction d'information, de résumé automatique de textes, d'organisation des connaissances à partir de documents de référence ainsi que de mémoires de traduction. Les applications de ces technologies concernent la veille stratégique, commerciale, scientifique et technique, la gestion des connaissances dans l'entreprise, la vidéo à la demande, la protection de la propriété intellectuelle, la formation par les technologies, la localisation de logiciels (adaptation linguistique et culturelle). Un laboratoire, le Lic2m (laboratoire d'ingénierie de la connaissance multimédia multilingue) dédié à ces activités a été créé au sein du List en 2002.



- **Le laboratoire d'ingénierie de la connaissance multimédia multilingue (Lic2m) du CEA List**

Le Lic2m a été créé au CEA List en février 2002. Il rassemble les compétences en ingénierie linguistique et en indexation d'images développées auparavant dans d'autres unités du CEA. Son activité est centrée sur la maîtrise de l'information exprimée sur différents médias (images fixes ou vidéo, texte, parole, musique, ...) pour produire des connaissances. Ce laboratoire est aussi spécialisé dans le traitement d'informations textuelles et parlées en différentes langues. Les équipes du Lic2m sont engagées dans plusieurs projets européens ou en partenariat avec des industriels. Les technologies développées font l'objet d'un transfert technologique à la société New Phenix.

CEA List

18, route du Panorama - BP 6 - 92265 Fontenay-aux-Roses Cedex
Tél. 01 46 54 99 37 - Fax. 01 46 54 99 42
www-drt.cea.fr



Le CEA

Un acteur clef de la recherche technologique

Acteur majeur en matière de recherche, de développement et d'innovation, le CEA intervient dans trois grands domaines : l'énergie, les technologies pour l'information et la santé et la Défense.

A travers la diversité de ses programmes, il poursuit deux objectifs majeurs : devenir le premier organisme européen de recherche technologique et garantir la pérennité de la dissuasion nucléaire.

Ses atouts pour y parvenir : une culture croisée ingénieurs-chercheurs, propice aux synergies entre recherche fondamentale et innovation technologique ; des installations exceptionnelles (supercalculateur, réacteurs de recherches, grands instruments de la physique, lasers de puissance...) ; enfin, une réelle implication dans le tissu industriel et économique.

Implanté sur 9 centres répartis dans toute la France, le CEA bénéficie d'une forte insertion régionale et de solides partenariats avec les autres organismes de recherche, collectivités locales et universités. Afin de favoriser le transfert des connaissances, il accorde une importance particulière à l'enseignement et à l'information du public.

Reconnu comme un expert dans ses domaines de compétences, le CEA est pleinement inséré dans l'espace européen de la recherche et exerce une présence croissante au niveau international.

Une variété de programmes articulés autour de trois grands axes :

Les énergies

Disposer de formes d'énergie compétitives, sûres et propres, en particulier non émettrices de gaz à effet de serre, tel est l'objectif que poursuit le CEA à travers ses recherches et développements sur l'énergie.

En appui des industriels, le CEA cherche à optimiser le parc actuel des réacteurs nucléaires et à mettre au point des solutions techniques pour la gestion des déchets radioactifs.

Il participe aux programmes de recherches internationaux sur les réacteurs et combustibles nucléaires du futur qui assureront une production à la fois plus économique, plus sûre et générant moins de déchets. Il conduit enfin des programmes sur l'impact sanitaire et environnemental de l'énergie nucléaire.

Les recherches du CEA soutiennent également l'essor des nouvelles technologies pour l'énergie : l'hydrogène, le photovoltaïque, la biomasse...

La fusion thermonucléaire, dont la maîtrise pourrait permettre dans l'avenir de disposer d'une source quasi infinie d'énergie, est également au cœur de ses recherches. Le CEA est ainsi fortement impliqué dans le projet international du réacteur expérimental ITER.

En amont des recherches et développements sur les énergies, il conduit différents programmes dans les domaines des sciences du climat et de l'environnement, des sciences

de la matière, de la chimie et des interactions rayonnement-matière.

Technologies pour l'information et la santé

Afin de favoriser l'innovation industrielle, le CEA dispose d'une recherche technologique de haut niveau dans le domaine des micro et nanotechnologies. Les applications industrielles de ces recherches concernent notamment les télécommunications et les objets communicants.

Il exerce ses compétences dans le domaine des technologies logicielles : systèmes embarqués et interactifs, capteurs et traitement du signal.

Grâce aux compétences qu'il développe autour des biotechnologies et des technologies nucléaires pour la santé (marquage biomoléculaire, imagerie médicale...), il est également un acteur de la recherche médicale.

Ces programmes appliqués s'appuient sur des recherches de base en nanophysique et ingénierie moléculaire, sciences des matériaux et cryotechnologies.

Au service de la Défense nationale

Dans le cadre des lois de programmation militaire, le CEA développe les programmes nécessaires pour garantir la pérennité de la dissuasion nucléaire française. A la suite de l'arrêt des essais nucléaires, il met en œuvre le programme Simulation, qui s'appuie sur d'importants moyens expérimentaux et de calcul (Airix, Laser Mégajoule, Supercalculateur Tera).

En matière de propulsion nucléaire (sous-marins, porte-avions), le CEA est notamment responsable de la conception et de la maintenance des réacteurs.

Il intervient enfin dans les instances nationales et internationales, où il contribue à la surveillance du respect des traités internationaux tels que le Traité d'interdiction complète des essais nucléaires (TICE). Il participe à la lutte contre la prolifération des armes nucléaires.