# *UN PROGRAMME DE RECHERCHE POUR L'ÉCONOMIE DE L'INFORMATION*[1]

Prof. Pierre Lévy, CRC, FRSC, Université d'Ottawa
1er sept 2006

**Résumé**
Les agents de l'économie de l'information sont des communautés de chercheurs qui alimentent et exploitent ce bien commun qu'est la mémoire numérique. Un des principaux problèmes des communautés de chercheurs est de disposer d'outils logico-symboliques leur permettant d'extraire le maximum d'information de la mémoire. Mon hypothèse est qu'un système d'adressage sémantique des concepts - conçu pour se prêter au traitement automatique plus facilement que les langues naturelles - ferait franchir un seuil décisif à la puissance des outils de recherche d'information. Pour traduire la diversité des significations exprimables en langues naturelles, ce système doit pouvoir adresser un espace conceptuel varié à l'infini. Pour satisfaire les plus exigeantes des communautés de chercheurs dans les sciences de l'homme et de la société, il doit autoriser le test de théories complexes sur la signification des documents, selon des méthodes reproductibles et avec des instruments de mesure mathématiques. L'hypothèse selon laquelle un système d'adressage universel des concepts répondant à ces contraintes multiplierait la puissance des outils de recherche d'information aujourd'hui disponibles peut être testée sur le métalangage de l'économie de l'information (IEML) décrit dans cet article et sur le site www.ieml.org.
Le programme de recherche que je propose s'articule en deux volets :
- premièrement, entreprendre la construction coopérative d'un moteur de recherche sémantique à sources ouvertes basé sur IEML ;
- deuxièmement, initier, au moyen de ce nouvel instrument d'observation, une exploration coordonnée de la mémoire numérique au service des sciences de l'homme et de la société.

**Abstract**
The agents of the information economy are communities of researchers feeding and exploiting a common good : the digital memory. One of the main problems of these communities of researchers is to get the best symbolic tools to extract a maximum of information from the memory. My hypothesis is that a system for the semantic addressing of concepts - designed for better automatic processing capabilities than the natural languages - would help the power of search tools to cross a decisive threshold. In order to translate the diversity of the significations expressible in natural languages, such a system should be able to address a conceptual space infinitely varied. In order to satisfy the most demanding of researchers communities in the humanities and social sciences, it should allow the test of complex theories about the meaning of documents, according to reproductible methods and with mathematical measurement instruments. The hypothesis according to which a universal addressing system for concepts fulfilling these constraints will multiply the power of the information search tools available today can be tested on the Information economy metalanguage (IEML) described in this article and on the website : *www.ieml.org*.

The related research program is articulated in two parts :
- first, undertaking the collaborative construction of an open-source semantic search engine based on IEML,
- second, thanks to this new observation instrument, initiating a coordinated exploration of the digital memory for the benefit of humanities and social sciences.

[1] NDLR : il s'agit ici de la première partie d'un article

# *UN PROGRAMME DE RECHERCHE POUR L'ECONOMIE DE L'INFORMATION*

**Introduction**

## Opacité de la mémoire numérique

Depuis l'apparition du Web au début des années 1990, les fondations techniques d'une économie globale de l'information ouverte et dynamique ont été posées. La mémoire numérique désormais accessible en ligne constitue le capital - ou le bien commun - de l'économie de l'information. Ce bien commun est alimenté par la création de documents numériques et il est exploité par des opérations de recherche : indexation des documents, formulation de requêtes et extraction d'information.

Or l'exploitation optimale du nouveau bien commun au bénéfice des communautés de chercheurs se heurte à d'importants obstacles dont les principaux sont:

- la fragmentation linguistique,
- l'incompatibilité mutuelle et l'inadaptation des nombreux systèmes d'indexation et de catalogage hérités de l'ère de l'imprimerie,
- les difficultés rencontrées par l'ingénierie informatique à prendre en compte la signification des documents au moyen de méthode générales,
- l'absence de transparence des méthodes employées par les moteurs de recherche commerciaux contemporains.

Deux grands programmes de recherche, le *Web 2* et le *Web sémantique*, tentent de répondre aujourd'hui, chacun à leur manière, au problème de l'opacité de la mémoire numérique.

## Le Web 2

Le projet du Web 2 est porté par une nébuleuse informelle de communautés qui s'activent principalement à multiplier les outils collaboratifs, bien souvent dans un cadre *open source* et P2P. Le Web 2 a tendance à considérer le Web comme une sorte de système d'exploitation pour des applications collaboratives en ligne. Cela se marque notamment par l'usage croissant des wikis, par la multiplication des processus de partage d'information tels qu'on peut notamment les expérimenter sur del.icio.us (partage de signets) et flickr.com (partage de photos) et par le succès des logiciels sociaux et des services tendant à accroître le capital social de leurs usagers (myplace.com est à cet égard emblématique).

Le succès mérité de Wikipedia, la vogue des modes de communication P2P, la montée continue des systèmes d'exploitation et des logiciels à sources ouvertes, la pression pour desserrer les freins que pose la propriété intellectuelle classique sur l'économie de l'information numérique peuvent également être considérés comme des tendances liées au Web 2.

Tout cela manifeste une exploration sociale des diverses formes d'intelligence collective rendues possibles par le Web et représente donc une évolution très positive. Mais, en fin de compte, il s'agit d'une exploitation par et pour le plus grand nombre de potentialités qui étaient techniquement et philosophiquement déjà présentes dès l'apparition du Web au début des années 90. Je vois dans le Web 2 une maturation culturelle et sociale du Web (qui a été conçu dès l'origine par Tim Berners Lee pour favoriser les processus collaboratifs) plutôt qu'un saut épistémologique majeur.

## Le Web sémantique

Quant au Web sémantique, contrairement à ce que laisse supposer son nom, il propose essentiellement des normes de codage *logique* des informations. Rejoignant certaines tendances du Web 2, l'ambition du Web sémantique est de constituer une sorte de système d'exploitation des données du Web au service des moteurs de recherche et des « agents intelligents ». Les principaux outils symboliques de cette

nouvelle couche du cyberespace sont :
- XML (*eXtended Mark-up Language*), dérivé du langage SGML de Charles Goldfarb, qui permet de décrire de manière universelle la structure des données ;
- RDF (*Ressource Description Framework*) qui permet de cataloguer les données du Web et le langage Sparkl qui permet d'interroger les ressources ainsi cataloguées ;
- OWL (*Ontology Web Language*), qui permet de décrire les « ontologies » c'est-à-dire la structure conceptuelle des divers domaines de connaissances.

Cet appareillage de descripteurs et de marqueurs a pour principale fonction de favoriser l'automatisation des traitements dans la recherche des données et l'exécution des opérations confiées aux agents intelligents ou robots logiciels.

**Le programme *open search***

Les deux orientations intellectuelles qui viennent d'être évoquées proposent des solutions certes utiles, mais partielles, aux difficultés de fond mentionnées plus haut. Le Web 2 définit plutôt un certain esprit, une orientation vers la croissance de l'intelligence collective. Le Web sémantique, pour sa part, se spécialise dans la définition consensuelle de normes favorisant l'interopérabilité en ligne.

Je propose ici un troisième programme de recherche au service de l'économie de l'information. Ce programme, baptisé *open search*, que je développe depuis 2002 au laboratoire d'intelligence collective de l'Université d'Ottawa, n'est nullement opposé, mais plutôt complémentaire à ceux du Web 2 et du Web sémantique. En effet, la réalisation d'un moteur de recherche ouvert capable de dissiper l'opacité sémantique ne peut que bénéficier aux outils collaboratifs - du côté du Web 2 - et aux normes d'inférences automatiques et de services informationnels - du côté du Web sémantique.

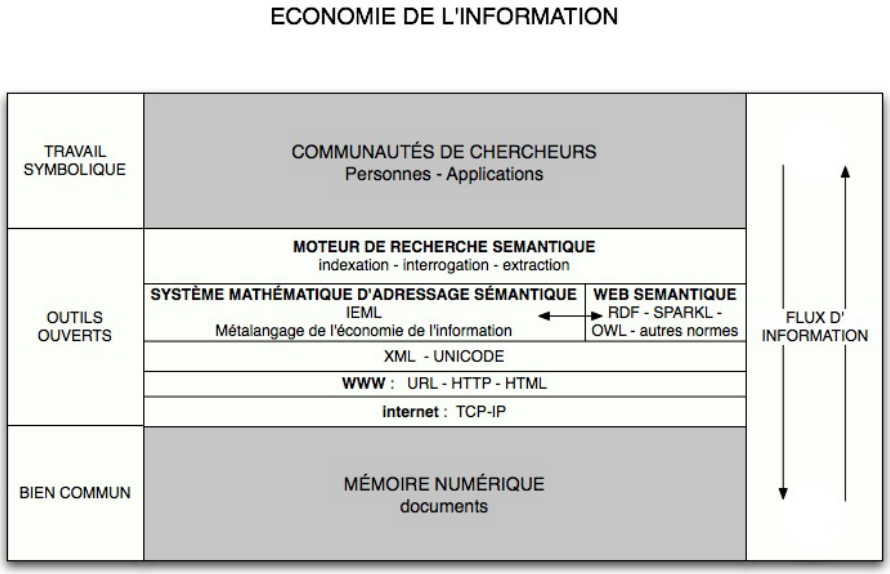
Le programme de recherche *open search* veut surmonter les problèmes auxquels est confronté l'économie de l'information en s'attaquant à leur cause : l'absence d'un *système d'adressage sémantique universel* (indépendant des langues et des cultures) capable d'optimiser la puissance et la portée de la recherche automatisée d'information. On peut supposer, en effet, qu'un tel système d'adressage, s'il était utilisé, résoudrait une grande partie des problèmes liés - je le répète - (1) à la fragmentation linguistique, (2) à l'incompatibilité des multiples systèmes d'indexation, (3) à leur inadéquation au traitement automatique à grande échelle, (4) à l'absence d'une approche systématique de la signification par l'ingénierie informatique contemporaine et (5) à l'opacité des méthodes et algorithmes utilisés par les moteurs de recherche commerciaux.

Une première version du système d'adressage sémantique dont a besoin l'économie d'information pour franchir un seuil décisif existe déjà : c'est IEML (pour *Information Economy Meta Language*). On en trouvera le noyau lexical et syntaxique sur le site www.ieml.org. Ce métalangage, développé au laboratoire d'intelligence collective de l'Université d'Ottawa[2], n'a pas vocation à devenir une langue parlée ou écrite d'usage courant au même titre que les langues naturelles comme le français, l'anglais ou le mandarin : ses fonctions sont d'indexer - ou d'adresser - les documents numériques rédigés en langue naturelle et de représenter des connaissances complexes à des fins de traitement automatique.

IEML (ou n'importe quel autre métalangage ayant les mêmes caractéristiques) peut jouer ce rôle de système d'adressage sémantique et permettre ainsi à l'économie de l'information de franchir les obstacles mentionnés plus haut parce qu'il réunit deux propriétés généralement séparées :

[2] Avec l'aide du programme des chaires de recherches du gouvernement fédéral canadien, de la fondation canadienne pour l'innovation (FCI) et du conseil de recherche en sciences humaines du Canada (CRSH).

- d'un côté, il est capable de traduire de manière distincte l'ensemble ouvert des *concepts* explicitables dans des langues naturelles ;
- d'un autre côté, contrairement aux langues naturelles, il peut être traité de manière optimale par les ordinateurs : c'est un système *mathématique* permettant des calculs puissants (mesures de distances sémantique, rangement sur critères sémantiques, inférences automatiques).




## Plan de l'article

Avant d'expliquer la solution que je propose au problème de *l'économie de l'information*, il est nécessaire d'en poser les termes le plus clairement possible. C'est pourquoi le premier chapitre définit les concepts principaux et les grandes fonctions de cette économie. Le second chapitre analyse les difficultés rencontrées aujourd'hui par la recherche d'information dans la mémoire numérique en ligne et esquisse les plans d'un *moteur de recherche sémantique* ouvert capable de résoudre ces difficultés. Comme le moteur de recherche sémantique ne peut fonctionner qu'au moyen d'un système d'adressage sémantique universel, le troisième chapitre décrit la *structure syntaxico-sémantique du métalangage* de l'économie de l'information. Ce chapitre explique comment IEML est capable d'adresser la signification des documents de façon mathématique tout en permettant la plus grande liberté et une variété virtuellement infinie dans l'expression des nuances sémantiques. Le quatrième chapitre, le plus long, analyse les possibilités de *recherche scientifique* dans la mémoire numérique offertes par le moteur de recherche sémantique. Ce quatrième chapitre est centré sur la description des *graphes conceptuels* IEML, qui peuvent exprimer aussi bien les données que les hypothèses au sujet des données.L'article se conclut par une discussion des thèmes du cerveau global et de la conscience réflexive de l'intelligence collective ainsi que par l'évocation de nouvelles perspectives de développement pour l'informatique et les sciences de l'homme.

### 1) L'économie de l'information

## Notions

Dans les réseaux numériques contemporains (le « cyberespace »), la *recherche d'information automatisée* met en rapport une multitude hétérogène *d'offres* d'information provenant de documents multimédia avec une multitude hétérogène de *demandes* d'information provenant des usagers des réseaux. Tel est

le cadre général de l'économie de l'information. Je propose d'appeler les agents de cette économie des *chercheurs* - plutôt que des usagers - afin de caractériser leur activité intellectuelle soutenue. Le chercheur (*homo intellectus*), qui travaille dans des univers de symboles numérisés, fait pendant à l'*homo economicus* idéal des économistes classiques.

On doit distinguer *l'économie de l'information*, qui concerne la création, l'échange et l'appropriation cognitive de l'information dans les réseaux techno-culturels et *l'économie basée sur la connaissance*, objet de la science économique classique, qui est plus particulièrement concernée par la valeur monétaire, directe ou indirecte, de l'information. Mais il est clair que ces deux économies sont étroitement interdépendantes, comme en témoigne le développement récent d'une *gestion des connaissances* qui prétend les articuler.

L'information de l'économie de l'information ne se réduit pas à la *face quantitative*, quasi matérielle, décrite scientifiquement par Shannon et Wiener dès les années 30 et 40 du XX° siècle sous la forme de jeux de différences dans des probabilités de combinaisons, écho de l'entropie physique dans des univers de symboles. Elle a également une *face qualitative* ou significative, porteuse d'une infinité potentielle de sens différents selon les situations, les contextes et les perspectives cognitives.

Les formes, les modalités et la puissance de la recherche d'information dans les réseaux numériques, et tout particulièrement dans l'Internet, représentent des facteurs conditionnants du marché global de l'information en voie de constitution. Or un tel marché favorisera d'autant mieux la production et l'appropriation de connaissances par *les communautés de chercheurs* (ou communautés virtuelles) qu'il sera ouvert et dynamique, par opposition à un marché fragmenté et à des échanges entravés.

De nombreuses collectivités économiques, sociales et professionnelles, ainsi que plusieurs gouvernements et organismes multinationaux, considèrent que des progrès scientifiques dans le domaine de la recherche d'information automatisée se répercuteront positivement sur la croissance d'une l'économie basée sur la connaissance et, ultimement, sur le *développement humain*.

**Analyse fonctionnelle de l'économie de l'information**

On peut décomposer le circuit complet de la recherche d'information en grandes *fonctions* dont chacune est susceptible de perfectionnement indépendant et dont la synergie avec les autres fonctions peut également être adaptée et améliorée en tenant compte des nouvelles possibilités techniques. L'analyse fonctionnelle qui suit considérer l'économie de l'information de façon abstraite, indépendamment des conditions matérielles et institutionnelles changeantes.

Je distingue quatre fonctions principales dans le circuit de la recherche d'information.

1) *L'indexation* produit des métadonnées explicites sur les documents pour faciliter l'extraction de l'information.
2) *La requête* explicite la demande des chercheurs d'information.
3) *L'extraction* sélectionne un corpus, analyse son contenu, propose une synthèse, range les données obtenues en fonction de leur pertinence par rapport à la requête et génère éventuellement des inférences à partir de la sélection, de l'analyse et de la synthèse.
4) *La production* des documents peut être un résultat direct de l'extraction ou peut en bénéficier indirectement, bouclant ainsi le cycle de la recherche.

Analysons maintenant plus en détail chacune de ces quatre fonctions.

**L'indexation des documents**

L'indexation d'un document produit un ensemble de métadonnées, c'est-à-dire des

données *au sujet* du document. J'utilise ici le mot « indexation » en un sens large qui comprend, selon les termes classiques de la bibliothéconomie, le catalogage (auteurs, titres, dates, éditions, etc.) et la description du contenu (thèmes, sujets, disciplines, mots-clés, tables des matières, index...).
Pour fixer les idées au moyen d'un exemple indépendant de l'existence des réseaux numériques, les *fichiers* et *catalogues* d'une médiathèque peuvent être considérés comme des *métadonnées* à l'échelle de l'ensemble des documents contenus par la médiathèque. Ces fichiers mettent en relation des auteurs, titres, média et sujets avec des « cotes » combinant les ordres alphabétiques et numériques. Les « fiches » concernant les documents peuvent être rangées par auteurs, titres et sujets tandis que les documents sont physiquement rangés par cotes. Au lieu d'attribuer de manière fixe tel document à telle étagère (comme cela se faisait encore au début du XIX° siècle), les cotes organisent les *positions relatives* des documents dans un ordre linéaire. Ces cotes représentent un premier niveau *d'adressage physique*, permettant de ranger et de retrouver les documents sur des étagères, tout en accommodant les transformations et croissances des bibliothèques.
Le premier niveau de métadonnées, *à l'échelle de la bibliothèque*, permet d'accéder à des documents. Mais le processus d'indexation ne s'arrête pas là. Il existe un deuxième niveau de métadonnées (tables des matières, index, glossaires, etc.) pour aider à l'extraction de l'information *à l'échelle du document*. Ce second niveau de métadonnées correspond à une échelle d'adressage physique de l'information par pages (et non plus par positions relatives sur des étagères). Il faut immédiatement noter que la situation des documents numériques dans les réseaux permet d'envisager des systèmes de métadonnées intégrés, cohérents et « fractals », c'est-à-dire sans séparation marquée entre différentes échelles de documents ou d'agrégation de la base documentaire. Le nouveau milieu numérique, par son ubiquité, permet également de détacher complètement l'adressage sémantique des nécessités de l'adressage physique des documents sur des étagères, dans des salles, etc.
Mais quelles que soient les différences de conditions techniques et institutionnelles, les diverses opérations de rangement, classement, adressage et description accomplie par l'indexation peuvent se ramener à une fonction principale de *production de métadonnées* sur lesquelles vont s'exercer des méthodes *d'extraction d'information*.
La production de métadonnées dépend elle-même de deux facteurs : le document lui-même et le métalangage d'indexation (par exemple : standards pour les entrées de catalogue, thésaurus index normalisant l'analyse et la classification du contenu). Les systèmes de catalogage et les métalangages d'indexation (terminologies, thesaurus, « ontologies ») sont d'autant plus puissants qu'ils autorisent des requêtes *variées* et des méthodes d'extraction *efficaces* sur une grande *quantité* de documents.

**La requête du chercheur**

La requête désigne une fonction d'explicitation de la demande d'information dans les termes d'un métalangage de requête. Le produit de cette fonction est un ensemble de métadonnées *sur la demande* capable de commander l'extraction d'information.
Lorsque nous faisons une recherche dans le catalogue d'une bibliothèque, nous devons exprimer notre demande d'information en termes d'auteurs, de titres et/ou de sujets listés par le (ou les) thésaurus utilisés par les bibliothécaires. Afin de pouvoir s'appliquer aux métadonnées qualifiant les documents, la requête doit s'exprimer - ou se traduire - précisément dans le métalangage qui a servi à les indexer. Dans un moteur de recherche commercial sur le Web, nous inscrivons un ou des mots

appartenant à une langue naturelle dans la zone de requête et le moteur ne va chercher *que* les documents qui contiennent ces mots dans la langue naturelle en question. Dans ce cas, le métalangage n'est autre que la langue naturelle commune utilisée par le document « *full text* » et le chercheur.
Comme on peut le constater, c'est donc le métalangage qui établit le medium symbolique commun à l'offre et à la demande d'information, dans lequel les processus d'extraction d'information (de l'offre vers la demande) vont pouvoir s'exercer.

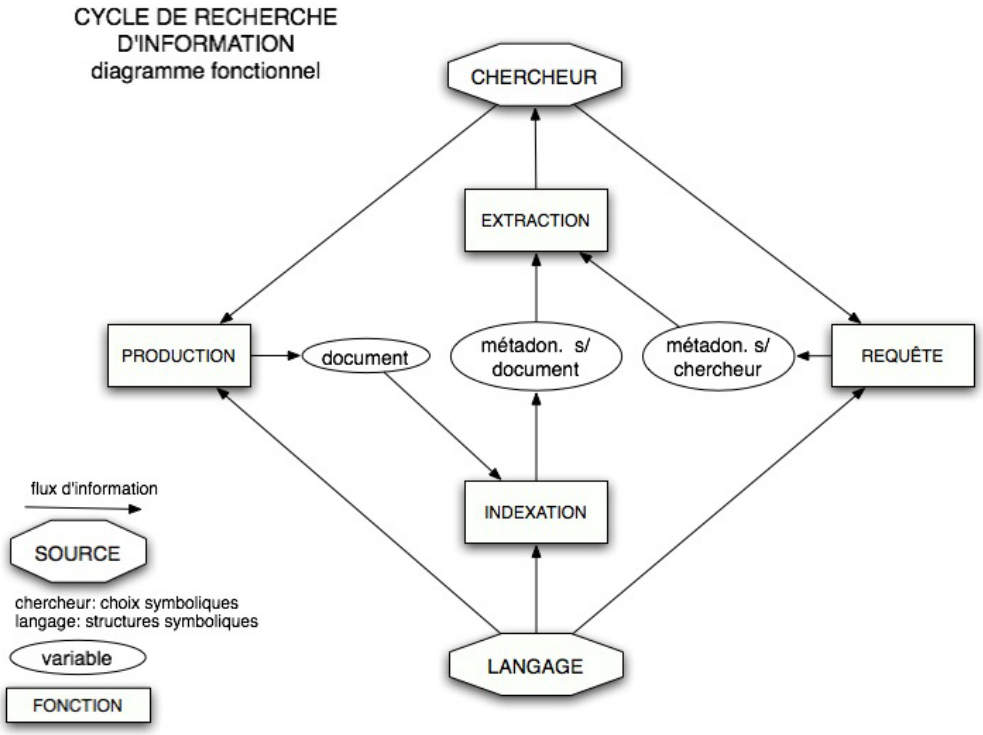
**L'extraction de l'information**

On vient de voir que l'indexation est une fonction de production de métadonnées dont les variables d'entrée sont le métalangage et l'offre d'information fournie par les documents. Parallèlement, la requête est une fonction de production de métadonnées dont les variables d'entrées sont le métalangage et la demande d'information du chercheur. L'extraction, finalement, met en rapport les métadonnées de la demande et celles de l'offre afin de produire une information destinée au chercheur.
Cette extraction peut être elle-même décomposée en plusieurs opérations interdépendantes.
- *La sélection* - ou le tri - d'un corpus pertinent à partir de l'ensemble de la base documentaire est la première étape de l'extraction. Les opérations suivantes vont s'exercer sur le corpus sélectionné.
- *L'analyse* consiste à décomposer le corpus sélectionné en ensembles de documents (ou parties de documents) distingués et regroupés en fonction de leur contenu.
- *La synthèse* abstrait d'une masse documentaire une information globale (holistique) compacte. La synthèse peut se représenter par un texte mais aussi par des images, graphiques et cartes de tous ordres, voire au moyen d'interfaces multimédia interactives.
- *L'inférence* produit de nouvelles informations en mettant en rapport des informations existantes et en leur appliquant des méthodes logiques de raisonnement. A la suite des efforts des ingénieurs en « intelligence artificielle » et des concepteurs de « systèmes à base de connaissance » de la deuxième moitié du XX° siècle, une des ambitions du Web sémantique contemporain est précisément de favoriser les raisonnements automatiques sur les informations présentes sur le Web au moyen de normes communes de représentation et d'indexation.
- *Le rangement* ordonne les informations obtenues en fonction de leur pertinence par rapport à la demande ou selon divers critères de mesure, d'antériorité ou de priorité. Le thème de l'ordre de rangement des réponses par les moteurs de recherche commerciaux sur le Web est devenu d'une grande importance économique et symbolique.

**La création des documents**

L'extraction d'information donne lieu *directement* à la production d'un document. C'est, si l'on veut, la « réponse » du système documentaire. Cette réponse est d'abord exprimé dans les termes du métalangage, qui code à la fois les métadonnées de la demande et celles de l'offre d'information. Par exemple, l'interrogation du fichier d'une bibliothèque fournit à l'usager la *cote* d'un livre à partir d'une requête de *titre*. Le titre et la cote sont sur la même fiche, et la fiche (la *réponse* du système) est elle-même un document. Des systèmes automatisés perfectionnés peuvent éventuellement « traduire » une réponse complexe exprimée en métalangage documentaire dans des formes mieux assimilables par le chercheur d'information.
Mais la recherche d'information contribue surtout à la création des documents de manière *indirecte*. Il devient de plus en plus facile de publier des documents sur l'Internet ou dans des « intranets » de communautés virtuelles de tous ordres

(entreprises, administrations, équipes de projets, associations, etc.). En outre, la navigation hypertextuelle (d'un lien à l'autre et de clic en clic) qui se déroule sur le Web donne au chercheur d'information le rôle d'un compositeur (ou d'un échantillonneur) de son document personnalisé. La distinction tranchée entre auteurs de documents et lecteurs passifs s'estompe donc au profit du rôle d'appropriation *et* de création d'information attribuée ici à un « chercheur ». Le gain d'information obtenu par le chercheur lui permettra peut-être de produire des documents contenant ou encapsulant de nouvelles informations, qui pourront à leur tour être extraites et réutilisés par d'autres chercheurs.




**Les sources de l'économie de l'information**

Comme on peut le voir dans le diagramme ci-dessus, les documents, métadonnées documentaires et métadonnées sur la demande d'information sont des *variables secondaires*, produites par les sous fonctions du cycle de recherche que sont la production, l'indexation et la requête. Mais si l'on considère le cycle de recherche comme une seule grande fonction - ou cycle complet - de l'économie de l'information, les variables primaires - ou *sources* - se réduisent à deux catégories : les chercheurs et les langages.

Les langages (y compris les langues naturelles et les métalangages documentaires) proposent des structures virtuelles permettant *l'articulation symbolique* de l'information. Les chercheurs actualisent ces structures symboliques par leurs choix d'offre (création) et de demande (requête) d'information. L'extraction d'information va « répondre » au chercheur à partir des métadonnées de l'offre et de la demande, en actualisant de manière pertinente la structure symbolique du métalangage documentaire.

Le diagramme fonctionnel du cycle de recherche d'information montre que le langage est une source directe de l'indexation, de la production de documents et de la requête. Il est aussi une source indirecte, mais déterminante, des processus d'extraction d'information. Quelles que soient les circonstances

techniques et institutionnelles, la cohérence et la portée du métalangage utilisé conditionnent la puissance des méthodes d'extraction.
Après avoir présenté les termes généraux de la fonction de recherche d'information, je voudrais maintenant examiner les problèmes posés à la recherche d'information contemporaine sur le Web.

**2) Le programme open search**

**La recherche opaque**

En 2006, la recherche automatisée d'information sur le Web est affectée de plusieurs graves défauts qui concourent à l'opacité de la mémoire numérique.

**Limite des méthodes utilisées par les moteurs de recherche**

Pour décrire le « contenu » d'un document, les moteurs de recherche commerciaux utilisent généralement des mots en langues naturelles. Mais (1) il existe des milliers de langues différentes, (2) à l'intérieur de chacune des langues, les mots peuvent avoir plusieurs sens et (3) le même sens peut s'exprimer par plusieurs mots, sans parler (4) des changements de sens dus aux variations de contextes et de points de vue.
Les algorithmes d'extraction utilisés par les moteurs de recherche contemporains travaillent sur des *chaînes de caractères* (en langues naturelles) et non pas sur des *concepts, thèmes ou notions*, qui sont en principe indépendants des langues *et* de leurs mots.
Sur Google, par exemple, les chercheurs expriment généralement leurs demandes d'information au moyen de mots dans une langue naturelle. Le vocabulaire de la langue naturelle en question fonctionne alors comme un métalangage de requête qui ne donne accès qu'aux documents indexés dans cette langue. A titre d'illustration, une recherche sur le mot « chien », *ne* donnera *pas* accès aux documents indexés par le mot « dog ». De plus, dans ce moteur de recherche, les mots des langues naturelles *ne* sont *pas* organisés en thésaurus ou en terminologies cohérentes comprenant des listes de synonymes. Pour garder le même exemple, une recherche portant le mot « chien » *ne* donne *pas* accès à un document indexé par le mot « canidé ».
Les moteurs de recherche commerciaux ne permettent pas non plus les recherches par auteur, titre et sujet. Il est très difficile, par exemple, de trouver un document sur le Web dont *l'auteur* soit George W. Bush. En effet, l'inscription de ce nom propre sur la zone de requête - même « avancée » - d'un moteur de recherche commande automatiquement comme réponse une immense liste de documents dont plus de 99% sont *au sujet de* George W. Bush.

**Limites des catalogues de médiathèques en ligne**

Dans le cas où le chercheur n'utilise pas un moteur de recherche commercial mais le catalogue d'une institution vouée à la conservation de documents (médiathèque, musée, etc.), la recherche pourra être beaucoup plus précise, bénéficiant d'un système de catalogage et d'indexation professionnel. En effet, les *langages documentaires* des bibliothécaires et des professionnels de l'information proposent des terminologies non ambiguës et bien structurées. En outre, leurs systèmes de catalogage permettent de distinguer les auteurs, sujet et éditeurs (par exemple) d'un document.
Mais, en contrepartie, la recherche ne pourra porter *que* sur les documents indexés par l'institution en question, et non pas sur tous les documents pertinents présents sur le Web. Chaque archive, bibliothèque, médiathèque ou musée en ligne utilise *un* métalangage d'indexation et *un* système d'extraction des informations, qui n'est pas forcément compatible avec les métalangages et les systèmes utilisés par les autres institutions. Cette séparation des documents en « silos » linguistiques et métalinguistiques limite fortement le bénéfice qui pourrait être

obtenu de la numérisation et de la mise en ligne des documents.
En outre, la plupart des *langages documentaires*, comme le « Dewey » ou la classification décimale universelle, proposent des hiérarchies de concepts ou de disciplines assez rigides, qui ne se prêtent pas de manière optimale au traitement automatique. La plupart des langages documentaires, même les plus souples - comme les langages à facettes inventés par Ranganathan - ont été conçus « avant les ordinateurs ».

**Limites des ontologies du Web sémantique**

Les *ontologies*, que les normes du Web sémantique recommandent de formaliser dans le langage OWL (Ontology Web Language) sont des réseaux sémantiques - le plus souvent des arbres ou des taxonomies - décrivant les relations entre concepts d'un domaine de connaissance. Or, d'une part, ces concepts sont exprimés par des mots en langues naturelles (avec tous les problèmes afférents déjà signalés plus haut) et, d'autre part, les ontologies - considérées comme structures de relations - *ne* sont *pas* traductibles les unes dans les autres. OWL permet seulement l'exécution d'inférences automatiques au sein d'une même ontologie. Cette fragmentation linguistique *et* logique des ontologies limite énormément les bénéfices potentiels du Web sémantique.
La même remarque peut être faite au sujet de RDF, dont la lacune tient à la formulation en langue naturelle du contenu des documents.

**Limites de la navigation hypertextuelle**

La possibilité d'une navigation hypertextuelle à grande échelle a été l'une des principales innovations apportées par le Web. Les hyperliens peuvent connecter diverses parties d'un même document ou relier d'un simple « clic » des documents différents, indépendamment de la localisation physique des serveurs qui abritent les documents connectés. Mais, là encore, toutes les opportunités ouvertes par la navigation hypertextuelle dans les réseaux numériques ne sont pas exploitées. En particulier, l'absence d'un adressage sémantique cohérent rend bien difficile la *génération automatique de liens*. Par exemple, même dans des corpus relativement homogènes, comme Wikipedia, fleuron du Web 2, aucun système de création automatique de liens ne connecte les articles portant sur les mêmes sujets ou sur des sujets complémentaires. Tous les liens doivent donc être créés « à la main ». La situation est encore pire si les documents sont rédigés dans des langues différentes.

**Absence de calculs de distance sémantique fiables**

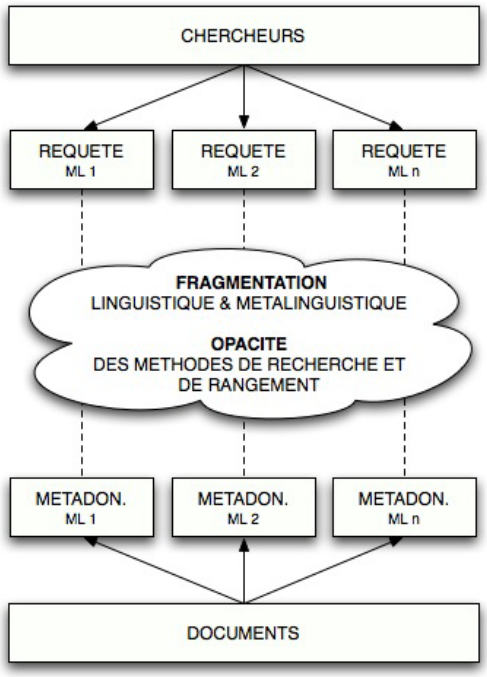
Faute d'adressage sémantique cohérent, ni les moteurs de recherche, ni les encyclopédies ou les bibliothèques en ligne n'autorisent de *calculs de distances sémantiques* un tant soit peu fiables qui permettraient, par exemple, d'aiguiller les chercheurs sur des informations « proches » des questions qu'ils ont posées si ces questions ne trouvent pas de correspondants exacts.

**Opacité des algorithmes de recherche et de classement des résultats**

Enfin, et ce n'est pas le moindre des facteurs qui font de l'Internet un labyrinthe, les moteurs de recherche commerciaux protègent comme des secrets industriels leurs systèmes d'indexation des documents et leurs algorithmes d'extraction d'information. Le chercheur est donc condamné à se contenter de résultats dont il ne peut pas savoir *exactement* selon quels principes et par quels moyens ils ont été obtenus. Le « rangement » des résultats sur les premières pages de réponses est à cet égard particulièrement problématique. Il est donc - en toute rigueur - impossible d'utiliser les moteurs de recherche commerciaux à des fins scientifiques. Il est également impossible au chercheur de personnaliser

telle ou telle méthode d'extraction d'information en utilisant ses propres critères explicites *d'attribution de crédit* (ou de degré de confiance) aux documents.

**ECONOMIE DE L'INFORMATION OPAQUE**




En somme, l'Internet a créé les conditions techniques d'un marché de l'information ouvert et transparent, mais l'absence d'un système d'adressage sémantique universel se prêtant aux traitements automatiques ne permet pas d'actualiser pleinement cette possibilité. On voit sur le schéma ci-dessus que la principale source de l'opacité de l'économie de l'information contemporaine est la multiplicité des métalangages (notés ML1, ML2, etc.) dans lesquels sont formulés les métadonnées sur les documents et les requêtes. La fragmentation linguistique et métalinguistique empêche les requêtes exprimées dans un langage ou dans une ontologie particulière de rejoindre les métadonnées exprimées dans un autre langage ou dans une autre ontologie. Pour ajouter encore à l'obscurité ambiante, les moteurs de recherche commerciaux dissimulent au public leurs méthodes d'indexation, de recherche et de rangement des résultats. C'est pourquoi l'économie de l'information reste encore dans le brouillard.

**Pour un moteur de recherche sémantique ouvert**

Le problème de l'économie de l'information, je le rappelle, est de perfectionner *le cycle de la recherche d'information* de manière à rendre la relation entre offres et demandes d'information la plus transparente possible. Or nous avons vu que les deux sources principales de ce cycle sont les langages et les chercheurs. Il ne peut être question, pour l'économie de l'information, de normaliser, de limiter ou d'instrumentaliser les choix des chercheurs. Je prétend, au contraire, qu'un perfectionnement du cycle de la recherche d'information doit permettre à l'offre de s'exprimer - et à la demande d'interroger - sur un mode autonome et de la manière la plus précise et la plus efficace possible. Puisque la liberté du chercheur est posée en principe, il ne reste donc comme possibilité pour résoudre le problème que d'agir dans la sphère du langage.

Mais comment agir ? La normalisation directe des langues et métalangages divers

dont se servent les chercheurs d'information serait une limitation de leur liberté, et cette voie autoritaire vient d'être abandonnée. C'est pourquoi j'estime que la solution du problème ne peut venir que de *la création d'un niveau d'abstraction supplémentaire* : un code d'adressage du sens - ou un système de coordonnées de l'espace sémantique - instaurant un réseau universel de passerelles automatiques entre les multiples langues et métalangages utilisés par les chercheurs. Ainsi, les langues et langages d'indexation actuellement en usage pourraient bénéficier d'une *correspondance* dans un métalangage universel qui offrirait en prime une puissance accrue d'extraction automatique de l'information. Le système d'adressage instauré par ce métaniveau pourrait être considéré comme *l'équivalent général* de l'économie de l'information, puisque toutes les demandes et toutes les offres pourraient s'y inscrire.
Cette hypothèse du système d'adressage sémantique peut être testée sur le métalangage de l'économie de l'information (IEML), développé au laboratoire d'intelligence collective de l'Université d'Ottawa. Comment fonctionne un moteur de recherche sémantique ouvert basé sur IEML ?

**Vers une méta-indexation**

Commençons d'abord par envisager la fonction d'indexation. Ici, deux possibilités principales se présentent. Ou bien les documents sont déjà indexés par un métalangage documentaire, ou bien ils ne sont pas.

Méthode métadonnées vers métadonnées IEML (M2M)

Si les documents sont déjà indexés de manière systématique, il suffit de traduire le métalangage d'indexation (thésaurus, ontologie, etc.) en IEML. Une fois le métalangage traduit en IEML, toutes les *métadonnées* sur les documents peuvent être *automatiquement* traduites en IEML.

Méthode documents vers métadonnées IEML (D2M)

Si les documents *ne* sont *pas* déjà indexés selon des méthodes systématiques, ce qui est le cas de la majorité des documents disponibles sur le Web, il faut utiliser des logiciels capables d'indexer automatiquement en IEML des documents « bruts » en langue naturelle, ou bien encore les *abstracts* ou séries de mots-clés qui sont censés décrire leur contenu de manière informelle. Pour ce faire, la méthode la plus économique consiste à (1) adapter des programmes d'indexation automatique en langue naturelle déjà disponibles (de tels programmes sont développés depuis plusieurs années par les spécialistes de traitement automatique du langage naturel) et à (2) leur adjoindre un module de *traduction de l'indexation* en IEML.

Méthode mixte

De nombreux cas sont intermédiaires entre les deux possibilités. Par exemple, le *learning object metadata* standard (LOM) propose un langage d'indexation très détaillé pour décrire les caractéristiques linguistiques, techniques, pédagogiques et juridico-économique (droit d'auteur) des documents mais laisse libre la description de leur contenu, ou de leur sujet. Dans de tels cas, on utilisera une *combinaison* des deux méthodes évoquées plus haut : M2M pour le catalogage systématique et D2M pour la description libre du sujet en langue naturelle.

Bénéfices de la méta-indexation en IEML

La traduction des langages documentaires et des ontologies en IEML aurait trois avantages directs.
- Premièrement, tout le travail d'indexation et de catalogage des documents qui est déjà réalisé serait utilisé et *sauvé* (il n'est pas à refaire).
- Deuxièmement, des documents rédigés dans des langues différentes ou des métadonnées qui ont été produites au départ en utilisant des métalangages

différents se trouveraient - à la fin du processus - exprimées dans un métalangage commun. Les ontologies et systèmes documentaires deviendraient donc mutuellement compatibles sur le plan logique. En particulier, des inférences automatiques et calculs de distances sémantiques pourraient être exécutées de manière *transversale* par rapport aux ontologies, terminologies et langues naturelles.

- Troisièmement, une fois traduite en IEML, une terminologie ou ontologie se trouverait automatiquement interprétée dans les langues naturelles en usage sur le Web. En effet, le dictionnaire multilingue IEML établit des correspondances automatiques entre mots ou groupes de mots en langues naturelles et *concepts IEML*. Grâce à ce dictionnaire, les *graphes conceptuels* IEML (textes IEML composées de concepts) sont éditables et lisibles dans toutes les langues naturelles supportées par le dictionnaire multilingue IEML.

**Des requêtes portant sur des concepts**

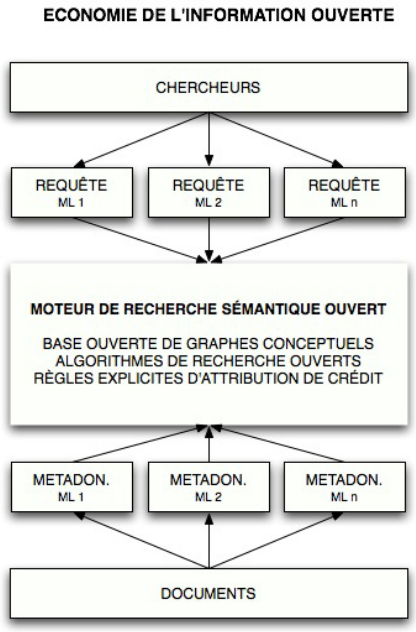
J'appelle *requête conceptuelle* la possibilité de faire des recherches non plus sur des chaînes de caractères mais sur des concepts, indépendamment des mots et langues différentes dans lesquelles les concepts sont exprimés. Aujourd'hui, ce type de requête n'est disponible que sur des corpus restreints et séparés, préparés par des spécialistes de la documentation automatique ou de l'intelligence artificielle. Un moteur de recherche ouvert pourvu d'un système d'adressage sémantique universel pourrait étendre progressivement la requête conceptuelle à l'ensemble du Web.

La généralisation de la requête conceptuelle suppose que les mots inscrits en langues naturelles dans le moteur de recherche sémantique puissent être automatiquement traduits en graphes conceptuels IEML et commander de ce fait des recherches portant sur *tous* les documents indexés en IEML, et non seulement sur ceux qui sont rédigés dans la langue naturelle utilisée par le chercheur d'information.

**Une production documentaire mieux connectée à la demande**

Une fois *l'offre* et la *demande* d'information exprimés dans le même métalangage, on peut disposer de données beaucoup plus fines qu'aujourd'hui sur leurs *rapports*. Sous l'angle de la production, on peut prévoir trois bénéfices principaux de la recherche sémantique ouverte.

- Premièrement, les documents répondront mieux aux demandes, puisque celles-ci seront mieux connues.
- Deuxièmement, on pourra mettre au point des méthodes d'indexation dynamique, capables d'optimiser la mise en valeur des documents produits et de s'adapter à l'évolution des requêtes.
- Troisièmement, on peut aussi envisager que des documents puissent être directement rédigés en graphes conceptuels IEML, notamment dans un contexte scientifique




## Une extraction d'information augmentée

Amélioration des méthodes actuelles

Comme je l'ai dit plus haut, les expressions du métalangage IEML sont des graphes de concepts IEML. Or chaque concept IEML peut être représentée par une chaîne de 1 à 20 caractères de l'alphabet latin sans accent, séparée par un ou deux blancs distinguant les mots-idées. Puisque *les graphes peuvent se ramener à des chaînes de caractères*, les algorithmes de recherche et d'analyse de l'information aujourd'hui en usage sur des textes en langues naturelles ou sur des listes de mots-clés pourront être *aussi* utilisés sur les graphes conceptuels IEML. La seule condition qui s'impose à ce réemploi est que les algorithmes ne soient pas spécifiquement liés à une langue naturelle ou à un système d'écriture particulier.

Or IEML est un système de notation idéographique et combinatoire. Dès lors, contrairement à ce qui est le cas pour les langues naturelles, les caractères et les différents degrés emboîtés de combinaisons de caractères sont *signifiants* et ils constituent en outre des *adresses sémantiques* uniques. Je fais donc l'hypothèse que les algorithmes de recherche et de rangement qui s'appliquent à des métadonnées exprimées sous forme de graphes conceptuels donneront des résultats plus riches et plus précis que s'ils s'appliquaient à des métadonnées exprimées en langue naturelle.

Extraction d'information fondées sur des méthodes originales

En plus des algorithmes de recherche déjà disponibles, dont l'efficacité est optimisée par le système idéographique et combinatoire du métalangage, il existe des *méthodes originales* de sélection, d'analyse, de synthèse et de rangement fondés sur les symétries et les niveaux d'articulation propres à IEML[3].

## Nouvelles possibilités de recherche et de navigation

L'indexation des documents au moyen de graphes conceptuels qui sont autant d'*adresses sémantiques* doit faciliter l'exécution des opérations suivantes :

- la génération automatique d'hyperliens entre documents ou parties de documents portant sur des sujets identiques ou complémentaires,

[3] Les principes généraux des méthodes fondées sur les niveaux d'articulation et symétries d'IEML sont exposés plus bas.

- le calcul de distances sémantiques, ou de degrés de similarité, entre métadonnées de requêtes et métadonnées de documents,
- la génération automatique de cartes sémantiques (synthèses) de grands corpus hétérogènes,
- la génération d'inférences et d'analyses automatiques portant sur des ensembles de documents « quelconques » sélectionnés par les utilisateurs selon leurs propres critères, même si ces documents ont été initialement produits dans des langues naturelles différentes et indexés par des métalangages différents.

**Transparence**

Par hypothèse, les métadonnées sur lesquelles travaillent les algorithmes d'extraction d'information du moteur de recherche ouvert sont toutes exprimées dans le même métalangage. Dès lors, il devient plus facile de *tester* et de *comparer* l'efficacité relative des algorithmes de recherche sur de grands corpus hétérogènes.

Le moteur de recherche sémantique ici préconisé doit de préférence utiliser des programmes informatiques *à sources ouvertes* pour ses opérations de sélection, d'analyse, de synthèse et de rangement. Le choix pour des programmes *open source* en concurrence garantit la participation d'une vaste communauté internationale de programmeurs à l'amélioration continue des logiciels et à l'innovation en la matière. Il permet surtout la transparence des processus de recherche et d'affichage des résultats. Contrairement à ce qui se passe aujourd'hui dans les moteurs de recherche commerciaux, on peut envisager que les chercheurs « avancés » puissent choisir en connaissance de cause les programmes d'extraction d'information qui conviennent le mieux à leurs finalités particulières, notamment en ce qui concerne les critères de confiance dans la valeur des documents (attribution de crédit) et les algorithmes de rangement des résultats.

**Des collaborations transversales simplifiées**

Quelles que soient les langues naturelles et les diverses ontologies ou langages spécialisés utilisés par le chercheur d'information pour formuler sa requête - d'une part - et quelles que soient les langues naturelles et les métalangages d'indexation ayant servi à produire initialement les métadonnées sur les documents - d'autre part - je répète que la traduction en IEML permet à une requête quelconque de déclencher automatiquement une extraction d'information portant sur *l'ensemble* des documents disponibles adressés par des graphes conceptuels. Dès lors, indépendamment du projet de moteur de recherche sémantique pour le public d'Internet, une indexation en IEML des *bases de données des entreprises* faciliterait les fusions et collaborations diverses, notamment dans un contexte multinational et multilinguistique. De la même manière, une indexation en IEML des *bases documentaires numérisées des agences, bureaux et ministères de différents niveaux de gouvernance* (municipal, régional, national et supranational) permettrait des échanges transversaux de données, la mise au point de tableaux de bords synthétiques pour l'aide à la décision, sans parler des nécessaires collaborations en situations d'urgence. Dans tous ces cas, l'indexation en IEML permettrait de plus un perfectionnement des méthodes d'extraction de l'information (sélection, analyse, synthèse, rangement).

Après avoir exposé, dans les lignes qui précèdent, les avantages que l'économie de l'information peut attendre d'un moteur de recherche sémantique ouvert basé sur IEML, je vais, dans le chapitre qui suit, entrer dans le détail des contraintes auxquelles le métalangage doit se soumettre pour remplir correctement son rôle, puis montrer comment sa structure permet de répondre à ces contraintes de manière satisfaisante.

**3) Le métalangage de l'économie de l'information**

## Le problème du métalangage

### Contraintes pesant sur le métalangage

IEML est une solution à un problème complexe d'ingénierie de la connaissance, ou d'architecture informationnelle. En effet, un métalangage pour l'économie de l'information doit satisfaire simultanément à trois contraintes : fonctionnelle, formelle et épistémologique.

*1) La contrainte fonctionnelle* est au service de *la liberté* d'expression des demandes et des offres d'information : elle exige de maximiser (1.1) la distinction effective des nuances sémantique et (1.2) les degrés de complexité sémantiques exprimables.

*2) La contrainte opérationnelle* est au service de *l'efficacité* : elle veut maximiser la puissance de la recherche automatisée d'informations (sélection, analyse, synthèse, rangement).

*3) La contrainte épistémologique* est au service de *la connaissance* scientifique : elle demande que la mémoire numérique affichée sur le Web soit justiciable d'une observation scientifique, permettant notamment de tester la pertinence de divers modèles et de valider ou d'invalider des théories.

Une liberté d'expression sans efficacité de la recherche, aussi bien qu'une efficacité du traitement automatique sans liberté d'expression du sens, seraient d'un bien faible secours pour l'économie de l'information. La difficulté vient de ce qu'il s'agit d'intensifier en proportion directe (dans la même direction) deux variables qui sont inversement proportionnelles pour le sens commun : la *variété des significations* exprimables par un système de symboles et la *puissance de la gestion* arithmétique et logique des symboles porteurs de significations. Quand à la contrainte épistémologique, seule sa satisfaction permettra à terme l'émergence et le développement d'une *réflexivité* de l'intelligence collective à partir de la mémoire numérique (je reviendrai plus longuement sur ce point dans la conclusion).

### La contrainte fonctionnelle : adresser un espace infini et symétrique

Pour répondre à la contrainte fonctionnelle, le métalangage doit envelopper virtuellement l'infini du sens et respecter l'égalité de principe et la symétrie entre les adresses sémantiques. Du côté de l'infini, il doit permettre un adressage distinct à tous les points de vue, idées et conceptions possibles dans l'univers de la cognition humaine. Du côté de la symétrie, un tel adressage ne doit refléter a priori *aucune hiérarchie particulière* entre philosophies, cultures, religions, langues, nations, professions, ministères, disciplines, théories, etc. Le métalangage répond donc à une sorte de métaphysique *pair à pair* (P2P) qui généralise *entre les concepts* une forme d'échange d'information déjà largement expérimentée dans le réseau *entre les serveurs*. Cependant, la stricte *équivalence* de leurs adresses cognitives ne préjuge en rien des intensités d'offre ou de demande qui s'attacheront aux graphes conceptuels dans différents « marchés de l'information », ou communautés virtuelles, libres d'établir leur propres règles d'échange et d'évaluation.

### La contrainte opérationnelle : augmenter la puissance de la recherche d'information automatisée

Pour répondre à la contrainte opérationnelle, le métalangage doit pouvoir être entièrement décrit comme une *grammaire formelle*. Cela signifie qu'il doit être décodable et manipulable par des méthodes logiques et mathématiques rigoureuses. La recherche et le traitement complexe des adresses cognitives représentant des significations distinctes doit pouvoir faire l'objet d'algorithmes entièrement explicites. Cette explicitation des algorithmes possède une dimension scientifique ou - si l'on veut - relève d'une

exigence de saine gestion des connaissances puisqu'elle favorise la transparence, le partage, les possibilités de tests, de comparaison, de perfectionnement, etc.
Cette explicitation possède aussi un aspect éminemment pratique puisque les algorithmes de manipulation des adresses cognitives sont destinés à être implémentés dans des programmes d'ordinateurs *efficaces* eu égard au traitement de l'information et *faciles à utiliser* de manière intuitive grâce à des interfaces utilisateurs multimédia appropriées. Le programme de recherche scientifique et technique qui se propose de développer ces logiciels *open search* relève plutôt de l'intelligence augmentée ou de l'intelligence collective que de l'intelligence artificielle au sens classique du terme. En effet, plutôt qu'une simulation directe de l'intelligence humaine individuelle, ces logiciels de manipulation de graphes conceptuels visent la compatibilité et la collaboration réciproque des opérations de recherche dans le cyberespace. Ils renforcent la coopération transversale et les capacités d'extraction d'information des explorateurs de données. Il ne s'agit donc pas d'intelligence artificielle au sens habituel. Néanmoins, comme on le verra par la suite, les logiciels composant le moteur de recherche sémantique peuvent bénéficier de nombreuses techniques de calcul et de représentation des connaissances expérimentées par l'intelligence artificielle depuis une cinquantaine d'années, tout en les complétant par le système symbolique de représentation du sens dont manquent ces techniques.

**La contrainte épistémologique : formaliser et tester des hypothèses scientifiques sur la mémoire numérique**

Le développement d'une recherche ouverte bénéficiera non seulement aux besoins d'information du public en général, mais aussi aux besoins - beaucoup plus exigeants, en ce domaine - de la communauté scientifique. L'adoption d'un système d'adressage sémantique universel (a) et l'explicitation des algorithmes de recherche d'information sur le Web (b) ont déjà été mentionnées comme conditions *sine qua non* d'une approche scientifique de la mémoire numérique. Mais cela n'est pas suffisant, car la satisfaction des deux conditions (a) et (b) doit ouvrir la voie à (c) : rendre possible la *représentation explicite* et le *test reproductible* de modèles et de théories concernant les données numériques. Cette possibilité de formuler des hypothèses testables concerne aussi bien l'ensemble de la mémoire déposée sur le Web que des sous-ensembles particuliers appartenant à des entreprises, gouvernements ou autres communautés virtuelles. Puisque l'objet de la recherche est la mémoire collective dans le cyberespace et que la plupart des grands musées et bibliothèques numérisent leurs fonds et les mettent en ligne, ce sont des modèles, théories et hypothèses relevant des *sciences humaines* qui nous concernent ici au premier chef, plutôt que ceux qui relèvent des sciences de la nature.
La satisfaction de la contrainte fonctionnelle (infinité virtuelle et symétrie d'un espace cognitif librement explorable) doit garantir que toutes les théories et tous les modèles puissent s'exprimer et qu'ils jouissent *a priori* d'un statut égal. Quant à la satisfaction de la contrainte opérationnelle (transparence et puissance des opérations) elle garantit une extraction d'information - une observation analytique des données - efficace et réalisée selon des procédures ouvertes. La satisfaction des deux premières contraintes concourt donc à la satisfaction de la troisième.

**Adresser un espace sémantique infini**

Le premier problème auquel est confronté la construction d'un moteur de recherche sémantique est de disposer d'un système symbolique qui réponde simultanément aux trois contraintes, fonctionnelles, opérationnelles et épistémologiques qui viennent d'être énoncés. Je vais maintenant

montrer comment la structure d'IEML est capable de répondre à ces contraintes.

**Sens et niveaux d'articulation**

La puissance de signification virtuellement infinie des langues naturelles est indissociable de leurs nombreux niveaux d'articulations : phonèmes, mots, propositions, phrases complexes, textes, contextes, etc. Pour atteindre une puissance équivalente, ou même supérieure - puisque augmentée par les capacités de mémoire et de calcul du cyberespace - le métalangage de l'économie de l'information s'articule, lui aussi, en multiples niveaux emboîtés.
Dans les langues naturelles, le sens ne réside pas dans un niveau d'articulation maître, qui s'imposerait aux autres niveaux. Au contraire, l'interprétation produit le sens par une mise en rapport dynamique des niveaux. Il en est de même pour IEML.
Puisque le métalangage de l'économie de l'information doit représenter le sens pour des ordinateurs et que les ordinateurs ne peuvent appréhender que des combinaisons de symboles explicitement réglées, IEML a été conçu comme (a) une idéographie (b) combinatoire.
(a) Qu'IEML soit une *idéographie* signifie que chaque symbole distinct, quel que soit son niveau d'articulation, doit avoir une signification distincte. Par contraste, on notera que, dans les langues naturelles, les phonèmes n'ont pas de signifié, que de nombreux mots de la même langue sont synonymes et que des phrases différentes peuvent avoir la même signification.
(b) Que cette idéographie soit *combinatoire* signifie que la signification d'une combinaison de symboles tend à correspondre à la combinaison des significations de ces symboles. Si ce dernier principe était appliqué à la lettre, on aboutirait à un langage trop redondant, à la couverture sémantique limitée. Le principe combinatoire est donc tempéré par un principe complémentaire *d'économie conceptuelle* selon lequel le maximum de « surface » sémantique doit être couverte par un minimum de symboles. Dans tous les cas, l'interprétation conventionnelle des symboles doit refléter autant que possible dans la sphère du « signifié » les *symétries* qui - comme on le verra plus bas - rendent les « signifiants » d'IEML éminemment manipulables par les ordinateurs.
L'interprétation des combinaisons symboliques d'IEML en langue naturelle fera l'objet d'une présentation spéciale[4].

**Eléments**

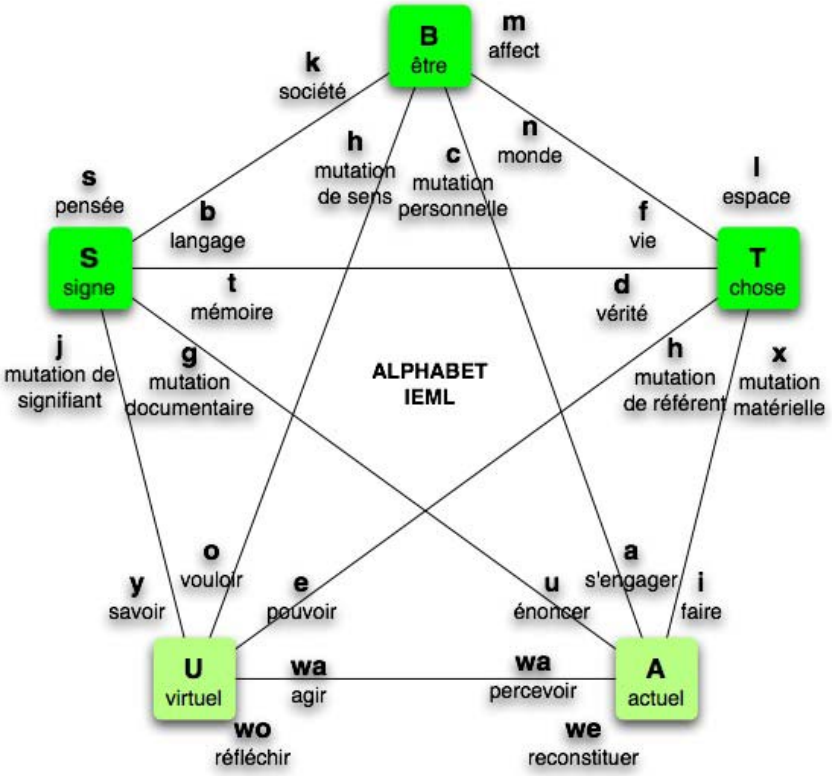
Le métalangage de l'économie de l'information repose sur des symboles élémentaires, de la combinaison desquels émerge la hiérarchie des niveaux d'articulation supérieurs.
Les symboles élémentaires d'IEML sont au nombre de cinq :
- Virtuel **(U)** et actuel **(A)** sont les deux éléments *pragmatiques*, liés à l'action, aux processus et aux verbes.
- Signe **(S)**, être **(E)** et chose **(T)** sont les trois éléments *sémantiques*, liés à la représentation, aux entités et aux noms.

**Flux d'information**

En IEML, toutes les combinaisons de symboles, quels que soient leurs niveaux d'articulation, ont la forme de *flux* d'information entre *stations*.
Ces flux d'information réunissent deux ou trois stations. La station source *envoie* le flux, la station destination *reçoit* le flux et la station traductrice *convoie* et transforme le flux entre la source et la destination. Le troisième rôle (traduction) est facultatif.
Les cinq éléments constituent les stations de premier niveau. Les stations de deuxième niveau sont constituées par les flux entre stations de premier niveau. Les stations de troisième niveau sont constitués par les flux entre stations de deuxième niveau, et ainsi de suite. C'est ainsi que toutes les combinaisons de symboles sont

[4] Au sujet des heuristiques d'interprétation en langues naturelles, voir en attendant le CI Lab technical report n° 1, qu'on trouvera sur le site www.ieml.org, à la rubrique « Journal ».

dérivées des cinq éléments de manière régulière.
Le schéma ci-dessous montre comment les 25 premiers flux (à traductrice vide) sont construits à partir des cinq stations élémentaires. Les éléments sont représentés par des lettres capitales et les flux entre éléments (ou *événements*) par des lettres minuscules ou des voyelles longues (wo, wa, wu, we). Les lettres minuscules sont placées du côté de leur élément source, à proximité de la ligne qui joint leur source à leur destination. Les lettres minuscules excentrées (s, m, l, wo, we) représentent des flux réflexifs (S vers S, B vers B...). Les 25 lettres minuscules de l'alphabet IEML vont à leur tour constituer des stations pour des flux (c'est-à-dire des combinaisons) de niveau d'articulation supérieur qui seront représentés comme des syllabes à 2 lettres, et ainsi de suite.




**Concepts**

A partir des éléments, IEML déploie quatre niveaux *finis* de combinaison et d'articulation qui sont appelés concepts. Les concepts IEML constituent les *caractères idéographiques* du métalangage.
- 25 ($5^2$) *événements* épuisent les flux d'information possibles entre éléments source et éléments destination.
- 625 ($25^2$) *relations* épuisent les flux d'information possibles entre événements source et événements destination. Les 625 relations ont été interprétées en langues naturelles[5].
- 240 millions ($625^3$ + $625^2$) *d'idées* épuisent les flux d'information possibles entre relations source, relations destination et d'éventuelles relations traductrice. Un peu plus de deux mille idées ont été interprétées en langue naturelle à l'été 2006[6].
- Une quantité astronomique de *phrases* ($10^{23}$) épuisent les flux d'information possibles entre idées source, idées

[5] Voir le dictionnaire IEML : www.ieml.org
[6] Voir le dictionnaire IEML : www.ieml.org

destination et d'éventuelles idées traductrice.
A titre de comparaison, il faut savoir que :
- le vocabulaire courant d'une langue naturelle comporte un peu plus de deux mille mots,
- il existe plus de cinquante mille idéogrammes chinois, qui ne sont maîtrisés dans leur totalité que par un petit nombre de lettrés,
- les plus grands dictionnaires d'une langue naturelle actuellement publiés comportent un maximum de 450 000 entrées[7].

**Graphes conceptuels**

Les concepts IEML sont en nombre astronomique mais fini. Ils peuvent être assemblés en une quantité virtuellement infinie, ou innombrable, de « textes » qui sont appelés en IEML des graphes conceptuels. Dans la version actuelle d'IEML, les graphes conceptuels peuvent prendre trois formes : ce sont des séries, des arbres ou des matrices régulières de concepts appelés claviers.
*Les concepts* IEML représentent l'ensemble des *caractères* possibles de ce système de notation conceptuelle qu'est le métalangage. Ils constituent un système d'adresses sémantiques *fixes*, intrinsèques au métalangage de l'économie de l'information.
*Les graphes conceptuels* organisent des flux d'information, plus ou moins complexes, *entre* les adresses fixes des concepts. L'adresse d'un graphe conceptuel est le graphe (au sens de la théorie mathématique des graphes) des adresses de ses concepts.
La possibilité *théorique* d'adresser un espace sémantique infini est donc assurée. Avec les moyens de traitement d'information et de communication contemporains, cette possibilité devient *pratique*.

**NDLR** : Fin de la première partie

[7] Comme, par exemple, le Webster Dictionary of American English. Le Grand Robert de la langue française ne compte que 100 000 entrées.