Contribution à la définition d'un vigiciel : quelle modélisation de l'information factuelle, événementielle et référentielle ?

Laure Berti-Équille, David Graveleau

Contribution à la définition d'un vigiciel : quelle modélisation de l'information factuelle, événementielle et référentielle ?

Laure Berti
GECT, Equipe Systèmes d'Information Multi-Média
Université de Toulon et du Var
B.P. 132, F-83957 La Garde cedex, FRANCE
berti@univ-tln.fr
tél. : (33) 04-94-14-22-20, fax : (33) 04-94-14-20-75

David Graveleau
Centre Technique des Systèmes Navals
DGA/DCE/CTSN/LSM/SF
B.P.28, F-83800 Toulon-Naval, France
gravel@newton.ctsn.dga.fr
tél. : (33) 04-94-16-26-52, fax : (33) 04-94-14-20-76

Résumé

Dans une perspective de conception d'un vigiciel ou système d'information intégré spécifiquement dédié à la veille technologique, notre réflexion s'attache à définir les besoins propres à ce type d'application en termes de contraintes liées aux données manipulées et aux utilisateurs concernés. Elle s'inscrit dans un thème émergent de recherche et de développement : la modélisation de la qualité d'une information par l'intégration de plus de sémantique dans la gestion des données formelles et informelles. Explorant les techniques actuelles de "découverte" et d'analyse d'information qui tendent à se spécialiser (recherche d'information IR, filtrage IF, extraction IE, gestion de données structurées / non structurées), on ébauche une évaluation de leur applicabilité au vigiciel. Dans ce cadre, notre article présente une analyse des besoins en terme de modélisation de l'information de veille, des meta-données associées et des différents modes de recherche correspondants.

Mots-clés

vigiciel, veille technologique, modélisation de l'information, qualité des données, recherche d'information,

---

Abstract

Our current work focuses on an emerging theme of Research and Development : the conception of Information Systems specifically meant for Technological Watch and Business Intelligence applications, we called Vigiware systems. Because the trend in the Databases and Information Systems community is to add more and more semantics to data inside their structure and format, the range of techniques for information searching is necessarly becoming more and more specialized : Information Retrieval IR, Information Filtering IF, Information Extraction IE, structured vs unstructured data. In this context, this paper presents a requirements analysis for specific information modelling and for information searching by various techniques, both necessary for an application domain such as Technological Watch.

Key-words

Vigiware, Technological Watch, Competitive Intelligence, Information Modelling, Data Quality, Information Retrieval

---

Contribution à la définition d'un vigiciel : quelle modélisation de l'information factuelle, événementielle et référentielle ?

1. Introduction 3
2. Nature des informations manipulées par le Vigiciel 4
3. Etat de l'art des techniques et outils disponibles 5
3.1 Recherche [IR], extraction [IE]et filtrage d'information [IF] 5
3.2 Systèmes de Gestion de Base de Données [SGBD] 6
4. Modélisation des informations manipulées par le Vigiciel 7
4.1 Typologie des sources d'information et de leur production 7
4.2 Tentative de réduction de l'information à des données : proposition 8
4.3 Tentative de réduction de l'information à des données : exemple 8
4.4 Application du Modèle et préservation de la dualité des modes de recherche 9
4.5 Coexistence des trois niveaux 10
4.6 Typologie des méta-données 11
5. Conclusions et perspectives 12
Références 12

1. Introduction

Pour contrer les menaces pesant sur sa position concurrentielle et les transformer en opportunités de développement commercial, une entreprise se doit de capter et d'exploiter, en temps opportun, les évolutions de son environnement : technologies émergentes, apparition de nouveaux produits, nouveaux fournisseurs, nouvelles " attentes " ou orientations du marché.

Sachant que, pour être efficace, une description opérationnelle de cet environnement technico-économique doit être adaptée aux missions et ambitions propres à chaque entreprise, partagée par tous ses collaborateurs et sans cesse alimentée par leurs actions sur le terrain, la mise en place d'un système d'information interne répertoriant ces entités (procédés, produits, clients, fournisseurs, ou plus généralement : technologies, systèmes et acteurs) et leurs multiples relations (dépôt de brevet, mise sur le marché, vente ou acquisition, alliance ou fusion et plus généralement toute relation acteur-système) est souvent envisagée. Outil d'aide à la décision stratégique pour les managers et outil de mémorisation voire d'analyse des informations collectées pour les veilleurs et les ingénieurs, le rôle central d'un vigiciel est alors décliné en 2 facettes complémentaires :

- diffuser aux acteurs l'information pertinente, systématiquement selon leur profil (*push*) ou à la demande (*pull*) pour la confection de dossiers techniques ou concurrentiels,
- mémoriser, classer et structurer les données au fil de leur récolte par les différents capteurs de l'entreprise (commerciaux, services marketing, de documentation et de veille, cellule brevet).

Mais, qu'il soit d'abord considéré comme un réceptacle des informations compilées ou d'abord comme support de communication d'une vision efficace de l'environnement, le système d'information pour l'intelligence économique ainsi défini doit cumuler deux capacités fondamentales :

- celle de stocker des informations multi-formes et multi-sources en entrée,
- celle de les restituer en sortie selon un format adapté à la nature des données et aux utilisateurs concernés.

Devant ainsi à tout instant faire coïncider les ressources disponibles aux besoins exprimés, il doit allier puissance d'expressivité (pour ne pas dénaturer l'information pertinente et correctement interpréter les requêtes) et capacité de normalisation, voire de traduction, toute information devant *in fine* être réductible à une forme assimilable par le système informatique.

Le problème de la modélisation de l'information au sein d'un vigiciel dans le but de promouvoir richesse des données et des traitements associés en vue de leur exploitation par tous les acteurs concernés comporte donc plusieurs aspects parfois contradictoires :

- incapacité du vigiciel, pour orientée que soit la veille stratégique ou la démarche d'intelligence économique, de restreindre a priori son champ d'intervention : aux informations formalisées (souvent essentiellement factuelles et numériques) qui constituent le noyau, on doit ajouter les informations informelles d'ordre événementiel concernant la vie des acteurs du marché et celles d'ordre référentiel permettant de juger de la fiabilité des précédentes et de les prolonger éventuellement par une recherche étendue.
- nécessité de traduire en données univoques, si possibles directement interprétables par l'utilisateur final, des informations par nature liées à un contexte de production particulier et dont les interprétations ou usages seront multiples selon le domaine d'intérêt ou l'expertise de chacun.
- difficulté de faire cohabiter au sein d'un même système une vision liée aux sources, ce que fait par défaut tout système d'information documentaire, et une vision liée aux entités manipulées (technologies, systèmes et acteurs), ce que fait une base de données structurées.

Quelle modélisation adopter pour préserver la qualité de l'information collectée, faciliter sa communication et son appréhension par les différents acteurs de l'entreprise et favoriser l'émergence d'un système d'information commun à la fois introspectif (critique par rapport à la fiabilité et la couverture de ses données) mais aussi entropique (favorisant spontanément son auto-enrichissement par la mise en corrélation de ses données et le test de leur originalité / cohérence), c'est cette réflexion que nous nous proposons d'ébaucher dans la suite de cet article. La section 2 tente de caractériser brièvement l'information manipulée par le vigiciel et introduit pour cela les trois niveaux actuel, événementiel et référentiel. La section 3 présente les techniques actuelles de recherche et filtrage d'information (IR - *Information Retrieval* / IF- *Information Filtering*), d'extraction d'information (IE - *Information Extraction*) et de gestion de données afin de mettre en lumière leurs différences d'approche et montrer que leur spécificités répondent à trois besoins particuliers existant dans le domaine de la veille. La section 4 décrira deux aspects essentiels pour la conception d'un système spécifiquement dédié à la veille : la modélisation de l'information et les différents modes de recherche. La section 5 conclut par les points qui feront l'objet de nos recherches et développements futurs.

2. Nature des informations manipulées par le Vigiciel

Même orienté correctement et en continu selon les besoins prédéfinis par la stratégie de l'entreprise, l'état des données déjà collectées et les nouvelles orientations du marché décelées au cours des analyses bibliométriques ou des synthèses sectorielles, un processus de veille est par nature universel et sans limite a priori : il fait feu de toute information nouvelle concernant son champ d'activité et doit cumuler les avantages de la détection et l'interprétation "à chaud" de

l'information critique et celle de la patiente recherche de "signaux faibles" noyés dans une accumulation de données individuellement peu signifiantes. Ainsi :

- si l'objectif est bien de surveiller les technologies émergentes susceptibles de toucher un domaine particulier, l'histoire des révolutions technologiques récentes démontre clairement le caractère transdisciplinaire de tout progrès décisif,

- si l'information pertinente est toute entière d'ordre technique ou technologique (les caractéristiques et les performances du système nouvellement développé concentrant l'essentiel du savoir utile), la difficulté d'obtention de ces données impose en pratique de :

  . bien sélectionner au préalable ses cibles (détection des systèmes intéressant en priorité l'entreprise),
  . connaître les acteurs du marché, notamment ceux dont la technologie est susceptible de dominer.

- si l'information accessible doit couvrir l'ensemble du spectre précédemment défini (technique, technologique, scientifique, économique voire sociale, et ce sur un grand nombre de secteurs), une recherche systématique pour multiplier ses sources d'information et retenir les plus pertinentes en fonction de critères de qualité évolutifs doit nécessairement accompagner en permanence la démarche globale. L'actualisation sera continue, éventuellement décentralisée et désynchronisée (au rythme imprédictible d'arrivée ou de prise en compte des informations nouvelles), pour un enrichissement progressif de la connaissance et une extension sans fin des domaines couverts.

- si la numérisation et la mise en réseau des informations collectées permettent d'envisager un service (dé)centralisé susceptible de fournir à tout instant une réponse adaptée au requérant (information d'ordre général ou question spécifique), quels que soient son niveau et le cheminement suivi par ses interrogations, il demeure la difficulté de faire cohabiter au sein d'un même système des informations de nature factuelle (description formalisée de systèmes et d'acteurs) et des informations de type événementiel (compte-rendus d'actualité mettant en scène l'histoire des relations entre acteurs et systèmes). A fiortiori, l'axe référentiel qui constitue à la fois le dossier justificatif des deux types précédents (Qui parle ?) en permettant au lecteur de se faire une première opinion sur la fiabilité des assertions et l'extension de celles-ci en lui permettant de prolonger sa recherche par l'exploration des autres facettes de cette source est le plus souvent complètement absent des systèmes d'informations techniques.

Ne pouvant se limiter arbitrairement ni sur le volume, ni sur la nature, ni sur la formalisation des informations à prendre en considération, le vigiciel doit pourtant être à même, pour remplir correctement sa mission, de traiter conjointement les trois niveaux d'informations définis ci-dessus : factuel, événementiel et référentiel.

- factuel : données descriptives d'entités objets de la veille, le plus souvent structurées voire exclusivement numériques ou codées, les informations de ce niveau requièrent des outils spécifiques pour leur saisie (convertisseurs d'unités, contrôles d'appartenance à un domaine de valeurs, méta-données indiquant la précision ou la nature), leur validation (cohérence et originalité par rapport à l'état des connaissances précédent) et leur restitution à l'utilisateur (masques de requète, filtres multi-critères, traitements statistiques de type OLAP - *Online Analysis Process*). Contrairement à une opinion couramment répandue, une valeur au sein d'un système d'information technique ne chasse pas forcément une autre, d'abord parce qu'il est quasiment impossible d'opérer la sélection en direct à l'entrée du système (sauf à disposer en permanence de tous les experts sectoriels) et d'autre part parce que l'incertitude fréquente régnant autour dela véritable nature de la valeur (origine, mode de mesure, précision, paramètre et version du système associé) interdit toute interprétation immédiate et définitive [Graveleau, 1997].

- événementiel : *news*, récits d'actualité mettant en scène les entités précédentes, ces informations, pour conjoncturelles qu'elles soient, ne peuvent faire l'économie d'une surveillance et d'une mémorisation continue. Peu formalisables par nature, leur lecture cursive et la consultation d'un historique significatif permettent d'orienter le recueil des données factuelles, modifiant les priorités accordées à chaque entité et réorientant le processus de veille. S'il paraît important de disposer d'outils spécifiques pour l'exploration de cette masse d'information, on cherchera du côté de la reconnaissance de concepts (noms communs) et d'entités (noms propres), à laquelle seront adjointes des fonctions de manipulation de graphes et de tris multi-critères (exploration des axes temporels notamment).

- référentiel : corollaire indispensable des deux niveaux arbitrairement séparés ci-dessus, l'axe référentiel apporte la crédibilité indissociable de toute fourniture d'information. Largement représenté dans le domaine documentaire par les méta-données bibliographiques, il est, curieusement, pratiquement absent des outils standard de gestion de données structurées de type SGBD, la complexité du réel à décrire étant artificiellement évacuée par l'hypothèse d'univocité (linéarité du point de vue du producteur de l'information à la donnée stockée, de cette dernière à l'utilisateur final du système).

Quel outils permettent de répondre à ce cahier des charges ambitieux et dans quelle mesure sont-ils spécifiques à un vigiciel, c'est ce que nous nous proposons de voir dans la suite de cet article.

3. Etat de l'art des techniques et outils disponibles

3.1 Recherche [IR], extraction [IE]et filtrage d'information [IF]

Après les nombreux travaux orientés sur la recherche d'information se résumant, le plus souvent, à de la recherche ou du filtrage documentaire multi-critère tels que les pratiquent les robots disponibles sur le Web (IR ou *Information Retrieval* et IF *Information Filtering*. cf tableau 1 en annexe), un nouveau champ s'est ouvert récemment avec l'extraction automatisée de l'information (IE ou *Information Extraction*). Née sous l'impulsion de projets à caractère militaire ou de Défense comme les conférences MUC (*Message Understanding Conference*) initialisées en 1989 par la DARPA [Irwin, 1995], cette discipline a pour ambition, à la croisée de plusieurs problématiques comme la traduction automatisée (MT ou *Message Translation*), le résumé, la synthèse automatique de textes ou la reformulation de questions ou de réponses, de dépasser la "simple" sélection de documents pertinents selon des critères liés au vocabulaire voire aux concepts pour extraire des "faits" de textes de forme et d'origine variées. Nouvelle formulation de problèmes fort anciens, elle est fortement influencée et favorisée par l'expansion rapide de l'information en ligne sur l'Internet et la relative faillite des moteurs de recherche actuellement mis à disposition pour la recherche d'information fortement contextualisée sur le Web. S'appuyant sur les nombreux travaux linguistiques de traitement

automatisé du langage naturel (TLN ou NLP, *Natural Language Processing*), son développement rapide s'axe sur une grande modularité des traitements successifs ou conjoints, l'analyse progressive de la morphologie vers la sémantique du texte se concentrant sur la structure et acceptant lacunes et erreurs [Pazienza, 1997] : dans une perspective où "quelques bonnes réponses valent mieux que rien du tout", l'objectif final de remplacement de l'opérateur humain (par exemple dans le cas du recensement systématique des naufrages à effectuer pour la LLyods à partir des coupures de presse internationales ou locales) est souvent remplacé par une spécification plus pragmatique telle l'aide au renseignement d'une base de données très structurée, telle celles compilant les actes terroristes (FBI) ou les organigrammes de société.

Dans ces derniers cas comme dans le nôtre, le problème se pose le problème de la formalisation dans des cadres prédéfinis d'éléments d'information ou de connaissance (templates) dans le but de la classer et de la mémoriser. Il apparaît cependant clairement que, si l'on peut imaginer sans grande difficulté une base structurée d'actes terroristes décrits par quelques mots clés circonstanciels (Quoi, Quand, Où, Qui,...), la description utile du champ complet couvert par les actions de veille stratégique à l'échelle d'une entreprise posent des problèmes d'un autre ordre de grandeur.

Machine à engranger et restituer des informations multiformes, le vigiciel ou système d'information à vocation "intelligence économique" d'une entreprise est bien souvent éclaté en un SGBD relationnel ou (orienté) objet pour le stockage des données, une base de documents et un ensemble de logiciels multi-usages pour les traitements statistiques (bibliométrie). Cette architecture hétérogène repose traditionnellement sur des logiciels de stockage de données formalisées adaptés aux types d'accès des veilleurs : documentaire ou table/fiche technique. Qu'on insiste plus sur la forme de la donnée ou sur le processus qu'elle subit ou suscite, qu'on décrive d'abord son domaine en termes d'entités ou de relations, qu'on y voit d'abord des propriétés à comparer ou mettre ponctuellement en commun ou un emboitement de hiérarchies multiples, enchevêtrement de points de vues dont la co-existence est légitime et ne menace pas l'équilibre de l'édifice, le choix d'une architecture d'un vigiciel intégré se pose d'abord en termes de gestion de données.

3.2 Systèmes de Gestion de Base de Données [SGBD]

Depuis les modèles de données traditionnels hiérarchique, réseau et relationel, de nombreux travaux se sont attachés à développer leur pouvoir d'expression en leur adjoignant de nouveaux concepts [Connolly, 1997] [Elmasri, 1994] [Ullman, 1988].

Dès lors que les structures de données proposées à l'utilisateur restaient très proches de celles utilisées dans la représentation physique, la sémantique des relations liant les données n'était pas explicitement précisée, ceci tranduisant immanquablement une perte d'information. Des tentatives de modélisation au plus proche du monde réel, classées parmi les "modèles de données sémantiques" furent alors développées afin d'offrir un plus haut niveau d'abstraction.

Les plus connus sont :

- le modèle Entité-Relation (ER) de Chen en 1976 [Chen, 1976],
- le modèle de données sémantique de Hammer et MacLeod en 1981,
- le modèle de données fonctionnel (FDM) de Shipman en 1981,
- le modèle association sémantique de Su en 1983

Leurs caractéristiques communes étaient d'offrir une représentation explicite des "objets" couramment manipulés par l'utilisateur, de leurs attributs et des relations qui les lient, notamment par des constructeurs de types abstraits (pour exprimer les relations de généralisation, d'agrégation, de groupement et d'instanciation). Les modèles de données sémantiques possèdent en conséquence trois qualités principales :

- celle d'accroître la séparation entre les niveaux conceptuel et physique, permettant ainsi une évolution fréquente du modèle conceptuel de données (MCD) pour une bonne adaptation aux changements du monde réel représenté dans le vigiciel,
- celle de diminuer fortement la surcharge sémantique des différents types de relations (un acteur du marché est multi-rôle),
- celle de fournir des outils d'abstraction efficaces pour manipuler les schémas (par exemple pour préciser le niveau de détail auquel on souhaite visualiser le schéma de la base), favorisant ainsi la multiplicité des points de vue.

Les modèles de données hyper-sémantiques ont ensuite été proposés afin de présenter de façon pertinente les situations du monde réel tout en étendant les capacités des modèles sémantiques par la gestion des connaissances à l'aide de concepts issus de la recherche en Intelligence Artificielle, tels que l'inférence, le flou ou les contraintes.

KDM (*Knowledge/Data Model*) de Potter et Kerschberg [Potter, 1989] permet de représenter de façon homogène à la fois la sémantique des données et celle des connaissances sous le paradigme représentationnel <attribut, objet, valeur>. Trois nouveaux constructeurs ont alors été ajoutés aux constructeurs classiques (tels que la généralisation, agrégation et groupement classification) :

- le constructeur de contrainte, plaçant une restriction sur un aspect d'un objet, opération ou relation avec d'autres objets / données,
- le constructeur heuristique, permettant d'affecter une relation de dérivation d'informations à des objets / données,
- le constructeur temporel, permettant de lier des types objet (types abstraits ou sous-types abstraits) au travers de relations synchrones ou asynchrones.

En réponse à la complexité croissante des applications en bases de données (CAO, Génie logiciel, gestion de données géographiques, gestion de documents, images, sons,...), deux "nouveaux" modèles sont nés : le modèle de données relationnel étendu et le modèle objet constituant la 3ième génération de SGBD.

Un objet peut être considéré comme une structure de donnée liée à une classe comme implémentation d'un concept, ou comme un agent en interaction avec d'autres agents [Gardarin, 1996]. Les modèles de données objet ont l'avantage vis à vis des modèles sémantiques d'apporter une plus grande souplesse :

- ils mettent l'accent sur le comportement des données en encapsulant dans le concept d'objet à la fois les données et les opérations possibles sur ces données,
- ils offrent un mécanisme d'héritage permettant de raffiner les classes en sous-classes plus spécifiques,
- la notion d'identité d'objet des systèmes orientés objet recouvre la notion d'entité des modèles sémantiques.

Ainsi, depuis une vingtaine d'années, les modèles de données tentent d'intégrer toujours plus de sémantique tout en répondant aux nouveaux types d'applications et aux nouvelles formes d'informations et de connaissances [Connolly, 1997] [Lebastard, 1993] [Elmasri, 1994] [Ullman, 1988].

Las, issus de conceptualisations successives, sans cesse perfectionnées pour y intégrer de plus en plus de complexité, les SGBD peinent à mettre leurs moyens au niveau des ambitions grandissantes des vigiciels. De simple gestion de fichiers plats aux derniers développements destinés à intégrer plus de réactivité dans leur modèle descriptif (agents, processus déductifs, ...), ils butent encore sur 2 obstacles fondamentaux, du point de vue du vigiciel.

- Non réductibilité de l'information de veille à une simple accumulation de données formalisées et structurées selon de simples hiérarchies,
- et son corrollaire, perte du contexte de production, de communication voire d'usage de la donnée une fois qu'elle a été extraite de la gangue documentaire qui colorait sa signification non univoque.

Ressortissant aux domaines de l'analyse sémantique voire philologique comme de la théorie de l'information, ces deux obstacles pour peu qu'on tente d'en épuiser les conséquences, ne peuvent être surmontés aisément, conduisant à des contradictions théoriques du type :

- l'information, même réduite à la relation d'un fait avéré, n'existe pas en soi ; elle ne peut être considérée que comme le produit intellectuel d'une opération complexe faisant intervenir, au minimum, un émetteur, un canal de communication et un récepteur visés [Shanon], chacun apportant, de par ses capacités et limitations culturelles ou physiques une orientation mettant à bas une quelconque neutralité ou pureté de la relation.
- sans support quel qu'il soit, documentaire ou électronique, l'information ne peut être appréhendée par un quelconque système d'information, aucune donnée n'acceptant longtemps une absence de transcription et donc de formalisation.

Concernant la mise en œuvre d'un vigiciel, ces paradoxes conduisent à des choix forcément réducteurs (mais) destinés cependant à garantir leur efficacité. On peut considérer désormais l'aspect modèle de données.

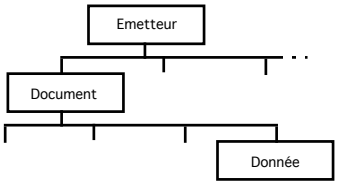
4. Modélisation des informations manipulées par le Vigiciel

4.1 Typologie des sources d'information et de leur production

Le projet de notre vigiciel étant motivé par la surabondance de données non concordantes et l'espoir d'être en mesure de qualifier chacune d'elle en vue de réduire l'incertitude globale et d'accroître et la fiabilité et l'étendue du champs de connaissance surveillé et maîtrisé, il est primordial de décrire au préalable les données et les sources accessibles.

Dans le continuum de la chaîne de production, de transmission et de réception qui brouille les étapes de transformation de l'information multiforme et les rôles des acteurs pour cette transformation, on réduira à trois le nombre d'entités canoniques : l'émetteur, sa production documentaire, les informations qui s'y trouvent.

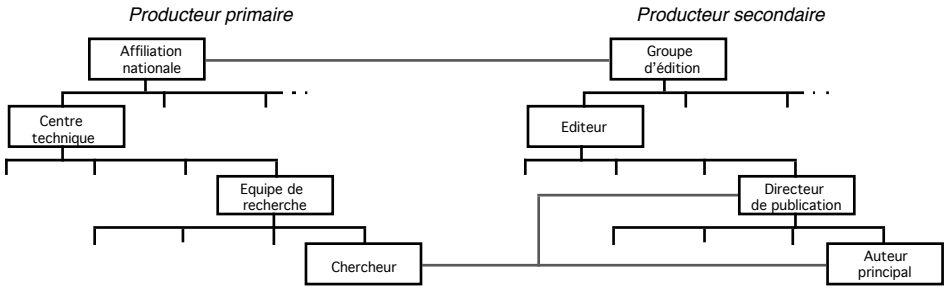
Typiquement, on considèrera donc une arborescence hiérarchique simple à trois niveaux :



Il est clair que cette formalisation est réductrice et qu'elle ne permet pas de transcrire adéquatement l'immense variété d'acteurs et des relations complexes qui les unissent dans la réalité : simple éditeur, directeur de publication, acteur principal, organisme d'appartenance de ces personnes (à quel niveau ? équipe de recherche, laboratoire, centre technique, affiliation nationale, pour ne citer qu'une chaîne classique dans la publication scientifique et technique).

De même, la notion de document peut prendre toutes les formes selon le support et le media considéré : ces formes multiples peuvent être considérées, selon les cas, à plusieurs niveaux (typiquement : collection, n° de livraison, page ou rubrique ou article pour les documents périodiques) et sont elles-mêmes unies par des relations multiples.

Enfin, les " informations qui s'y trouvent " est un vocable pratique pour parler des données que le vigiciel a pour mission de stocker et diffuser ; il masque lui aussi une grande complexité tant il est vrai qu'on ne peut simplement réduire une information à une accumulation de données complémentaires, le regard de l'utilisateur opérant un filtrage souvent bien différent de celui opéré par le producteur initial.



Si nos trois niveaux de modélisation des entités liées à la chaîne informative sont donc imparfaits, ils laissent au veilleur la responsabilité de décider du niveau de représentation (granularité) et de la richesse des liens à représenter. Ce choix effectué au cas par cas permet une certaine optimisation du rapport effort de mise en forme/efficacité auprès des utilisateurs.

4.2 Tentative de réduction de l'information à des données : proposition

La sphère médiatique est bruissante de mots ; relations de faits, interprétations et expression d'opinion s'interpénètrent au point d'interdire une objectivité idéale car dénuée de point de vue particulier. A l'opposé, ce qui vaut pour le producteur d'une information et son dernier médiateur connu reste valable pour le futur utilisateur du vigiciel, dont les intérêts et la culture guident la recherche, orientent le regard et le jugement sur la pertinence et la fiabilité.

Devant l'impossibilité d'extraire toutes les données et d'exprimer sous forme de méta-données toute la richesse du contexte de l'information considérée, on ne peut que proposer une voie médiane, adaptée nous semble-t'il aux contraintes informatiques et aux besoins de l'utilisateur.

Elle se fonde sur le double principe suivant :

- extraction, habillage et stockage de la donnée utile sous une forme compatible avec le modèle conceptuel des données
- conservation du lien de la donnée avec sa source, stocké sous la forme d'un signalement documentaire éventuellement d'un extrait ou de tout le document considéré.

Le premier aspect possède le triple avantage de :

- présenter un niveau d'abstraction et de formalisation suffisant pour permettre des comparaisons formelles voire des calculs sur des données similaires d'entités apparentées,
- situer simplement la donnée dans les arborescences des entités et des données de la base, ce qui autorise une navigation experte,
- permettre d'exprimer à la fois le point de vue du veilleur sur le contexte d'émission ou de recueil de la donnée et celui de l'expert sur son usage potentiel et ses limites. On retrouvera ainsi, par exemple, un avis critique sur la rareté ou la nouveauté de la publication et sur la fiabilité supposée.

Le second vise à pallier les inconvénients de l'opération d'extraction-traduction-réduction de la donnée en donnant à l'utilisateur final la possibilité de revenir " à la source " et de juger par lui-même, une fois la donnée réinsérée dans son contexte (et enrichie de méta-données signalétiques), la pertinence de son point de vue sur celle-ci.

C'est à la fois un remède aux erreurs de manipulation (saisie ou conversion erronée) ou aux lacunes du processus. Typiquement, une donnée factuelle représentée par un paramètre et une valeur numérique prête le flanc à d'innombrables simplifications réductrices : traduit-elle bien la précision de la mesure (" de l'ordre de 100 m "), la certitude du fait ou de l'avis exprimé (" pourrait comporter 1200 capteurs "), sans compter l'ambiguité sur la cible (quelle version du système est décrite ?), la nouveauté ou la rareté relative du fait relaté (scoop, donnée volatile, conjoncturelle ou structurelle) voire sa distance à la vérité (hypothèse, fait avéré, observation, ...) ?

4.3 Tentative de réduction de l'information à des données : exemple

Si l'on examine maintenant chacune de ces approches selon le type d'information et les outils de stockage et de recherche qui correspondent, on peut considérer les formes archétypales suivantes :

- information d'actualité : 1 récit d'un fait daté et situé par une source identifiée

```
3 mars 1998 : La société Thomson Marconi Sonar annonce le succès des essais à
la mer de son nouveau sonar actif très basse fréquence CAPTAS 20 vendu à
l'exportation aux Emirats ARabes Unis....
```

- Donnée descriptive : 1 fiche technique (datée) d'un système établie par une source identifiée

```
nom du sonar        :     CAPTAS 20
Type                :     actif
Gamme de féquence   :     TBF
...
```

Ce qui donne dans le vigiciel,

info 1 : résultat d'essais d'un système " S " par l'acteur-fabricant " F " à la date " j ", au lieu " l " et dans le cadre du programme " P ".

info 2 : export d'un système " S " par l'acteur-fournisseur " F " à la date " j ", au lieu " l " et au profit de l'acteur-client " C "

info 3 : description d'un système " S " par l'acteur-fournisseur " F "

ce qui donne au niveau explicite, une action mettant en relation un objet incomplètement décrit et des acteurs dans des circonstances plus ou moins définies. Le premier travail du veilleur consistera à clairement identifier voire reconstituer " S ", " F ", " C ", et éventuellement " l " et " j ". Puis il explicitera le contexte de production-recueil du fait : le fait est raconté par l'acteur-media " M ", sur les indications de " F " ou " C " (annonce à visée promotionnelle), dans un document " D ", publié par " M " à la date " j+n ", et destiné à un public " P ".

Le veilleur effectuera son travail de signalement et de caractérisation du contexte, notamment pour expliciter autant que faire se peut les intérêts en jeu et leur influence sur la fiabilité des données publiées (quels acteurs derrière " M " ? Pourquoi et pour qui publier ?). L'analyse fine de ces dernières doit en revanche clairement être dévolue à l'approche technique.
On voit l'importance de la dimension référentielle, produit du processus " veille technologique ", qui :

- établit et met à jour la liste des systèmes " S ", des acteurs " F" et " C ", et des relations qui les unissent dans le temps. On retrouve typiquement des informations du type :

  j : F-C(S) vente d'un système
  j : F-F(S1...Sn) association ou fusion
  j : F(S) conception / mise sur le marché d'un système.

- concentre son attention sur l'approche documentaire (Qui parle ? pour qui ?), ce qui donne :

  [j:F(C)]M... où "M" est plus ou moins proche et dépendant de "C" ou "F"

Ce travail, axé sur l'originalité de la publication, permet la constitution et la mise à jour d'un état de la connaissance. Son résultat comporte deux types de sortie exploitables par le spécialiste et le manager, dans des visions complémentaires :

- organisation de la masse documentaire pertinente, permettant au spécialiste de revenir à la source (pour consolider son opinion sur une donnée),
- mise en perspective des acteurs et des systèmes, pour in fine, établir un classement actuel voire prospectif des " gagnants " et des " perdants " (par la vision des media, *ie* le filtre de la notoriété).

4.4 Application du Modèle et préservation de la dualité des modes de recherche

Une modélisation duale de ce type permet in fine de concilier les 2 modes de recherche traditionnellement proposés aux utilisateurs :

- l'approche documentaire, caractérisée par une recherche symbolique sur des textes peu formalisés décrivant des faits (événements, récits, opinions, interprétations)
- l'approche technicienne, caractérisée par une recherche de valeurs numériques de paramètres sur des fiches structurées décrivant des objets réels.

		Point de vue				
	Approche	*Accès par...*	*Information recherchée*	*Acteurs*	*Relations*	*Processus central*
Veilleur	*Documentaire*	**Source**	**Actualité**	Emetteur & media		**Collecte**
		• référence bibliographique	Fait daté	• producteur I	• acteur-système	
		• thème	&	• producteur II	• acteur-acteur	
		• système	situé		• système-système	
Spécialiste	*Technique*	**Système**	**Fiche technique**			**Description**
		• nom, concepteur		Concepteur	• paramètre	
		• thème	Valeurs de paramètre	Fournisseur	système	
		• référence bibliographique		Client	• système-système	

• La première approche correspond idéalement à la vision documentaire, associant le point de vue des sources (ou niveau " acteurs ") et celui de l'actualité (au niveau des " données recherchées "), et centrée sur les opérations de collecte de documents (au niveau des " processus ").

Il est clair cependant qu'une telle modélisation a aussi l'avantage de promouvoir la mise en commun de points de vue et d'outils, quand cela apparaît naturellement : on peut citer la recherche par thème qui regroupe différentes coupes transversales (par famille ou nationalité de système ; selon l'axe du temps en se concentrant sur les évolutions des offres ou des demandes). De même à la frontière de chacune des 2 approches canoniques, le veilleur comme le spécialiste peut s'aventurer sur le terrain de l'autre . On retrouve ce type de comportement lorsque :

- le technicien met en question la fiabilité d'une description technique proposée (Quelle source ?) et désire juger par lui même selon son appréciation du document ou de l'émetteur,
- le veilleur doit orienter son processus de collecte (Quels paramètres prioritaires sont mal connus ?).

• Produit du processus " expertise de données ", la dimension technicienne est, quant à elle, implémentée dans la chaîne d'analyse et de synthèse " extraction / formalisation / traduction-validation / restitution-recommandation " où la préservation de la traçabilité entre les étapes de traitement maintient le lien avec la dimension référentielle évoquée ci-dessus.

En effet, la mise en œuvre successive de ces 3 niveaux d'analyse et de synthèse, avec la traçabilité qui tient lieu de dossier de justification minimal, est garante, à notre avis, de l'amélioration progressive de la connaissance dans le respect des contraintes contradictoires liées aux données et aux utilisateurs :

- sa lourdeur relative est contrebalancée par la mise en facteur des efforts effectués pour des utilisateurs dont les niveaux d'information et les thèmes sont multiples et évolutifs.
- le recueil de l'information par le processus " veille stratégique " est forcément désynchronisé par rapport aux besoins exprimés par les (groupes d')utilisateurs sur (l'ensemble des) thèmes retenus : il suit une logique propre aux données, celle de leur accessibilité.
- l'appréciation de la qualité d'une donnée évolue elle aussi, selon le regard de l'utilisateur et l'ensemble de la connaissance acquise.

4.5 Coexistence des trois niveaux

Le modèle multi-niveaux de la dimension factuelle du vigiciel se traduit ainsi :

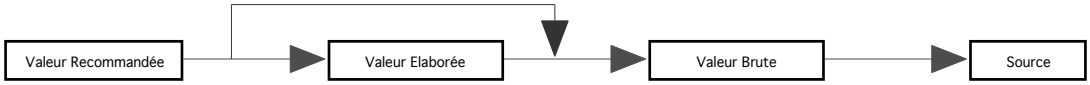
Ressource	Etapes du processus		Typologie des données		
Publications »	Extraction-traduction	Appréciation	Notation	Valeurs Brutes	Traçabilité
Modèles de cohérence »	Validation - restitution	Appréciation	Notation	Valeurs Elaborées	Traçabilité
Algorithmes de sélection multi-critères »	Recommandation	Appréciation	Notation	Valeurs de Référence	Traçabilité

Les outils de rétro-action pour le contrôle de la validité du processus global correspond ainsi à la confection d'un dossier de justification comportant :

• l'origine de la valeur recommandée (valeur brute et/ou valeur élaborée)
• la pertinence des traitements (critères de sélection et d'appréciation, modèles de validation ou d'élaboration).

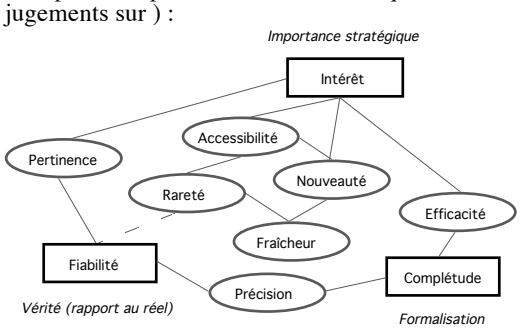
Ce modèle induit naturellement que l'on habille progressivement la donnée à l'aide de méta-données continuellement enrichies, la perte représentée par l'extraction de sa gangue documentaire porteuse de son contexte de production-émission devant être :

- compensée par l'explicitation des indices de performances sur la grille de critères élaborée précédemment (nouveauté/rareté/pertinence/efficacité...), ces indices représentent un premier niveau de méta-données, le second (la notation/cotation) concentrant le jugement de fiabilité émis en parallèle
- passagère et réversible, le lien référentiel permettant, selon le principe de traçabilité, un retour vers les niveaux inférieurs :



4.6 Typologie des méta-données

La liste des méta-données susceptible d'être établie et implémentée dans le modèle semble à la fois inépuisable, fortement intriqué et très dépendante des types de données et des utilisateurs potentiels. Leur échelle d'appréciation, leur poids respectif et les relations qui les unissent sont eux-même sujets à caution. On citera pour exemple (les jugements sur) :



Etablir une carte exhaustive et universelle de ces critères de qualité et des indices qui leurs sont liés semble une gageure. Tout au plus pourra-t-on proposer quelques oppositions ou rapprochements heuristiques :

Qualité : intérêt, fiabilité, complétude où :
- intérêt : fort, moyen faible, nul, non apprécié
- fiabilité : sûr, probable, possible, douteux, faux, non appréciée
- complétude : totale, partielle, laucnaire, non appréciee

Nouveauté : accessibilité, rareté, fraîcheur où :
- accessibilité : constante, périodique, exceptionnelle
- rareté : commune, expert, exceptionnelle
- fraîcheur : mois courant, année courante, année pénultienne, ...

De même, les jugements suivants s'opposent généralement :

rareté <-> fiabilité
efficacité/maniabilité <-> complétude
efficacité/précision

Le choix optimal doit bien entendu être fait en fonction : du type du volume et de l'importance stratégique des données et, in fine, des efforts susceptibles d'être consacrés à la justification des données de référence.

En pratique, notre modèle à trois niveaux repose sur 2 concepts fondamentaux :

- la cotation, indice de qualité tourné vers l'appréciation progressive (par notation) de la fiabilité
- la caractérisation peu formalisée et adaptée à chaque (type de) donnée.

Cette dernière retiendra donc le trait saillant et fortement contextualisé de la donnée vue à travers le filtre du système d'informations : l'indice sera donc un jugement bref sur (selon les cas) son originalité :

. la (faible/grande) précision (inhabituelle) liée au capteur (hypothèse, mesure, observations,...)
. la nouveauté, rareté ou le caractère exceptionnel de la publication par ce type de source
. la concordance/contradiction avec le savoir antérieur ou les autres données da ls source
. l'importance stratégique (exceptionnelle).

5. Conclusions et perspectives

Dans une précédente contribution [Graveleau, 1997], nous avions exposé une démarche centrée sur l'organisation d'une collaboration étroite entre le veilleur et l'expert pour la validation de l'information factuelle et la justification de la fiabilité de la valeur finalement recommandée entre toutes les données disponibles par le référencement de son origine et des traitements de restitution liés à l'application de règles de cohérence (pertinence / vraisemblance). Nous proposions la mise en place d'un système informatisé de Veille constitué autour d'une base multi-sources et de modules de criblage. Outil de dialogue entre les acteurs, son objectif primaire était de permettre la fourniture, à tout instant, d'une valeur recommandée (la meilleure valeur disponible) susceptible d'être accompagnée d'un dossier justificatif mentionnant son origine, les valeurs concurrentes éliminées et son degré de fiabilité supposé (cotation).

La prise en compte plus étroite du contexte de production, de communication et d'utilisation de l'information d'intérêt technologique impose désormais d'élargir le champ d'activité de ce type de système en intégrant les données peu fomalisées, notamment celles de type événementiel (*ie*, où les composantes temporelle et conjoncturelle / relationnelle dominent l'aspect purement descriptif). Cette extension, en conservant l'axe référentiel comme garant ultime de la crédibilité de la donnée recommandée (dossier justificatif) pose cependant le problème de la modélisation conjointe de ces trois niveaux d'information (factuel, événementiel et référentiel) et des outils de manipulations spécifiques qui peuvent les exploiter.

Dans une perspective d'approfondissement des mécanismes d'élaboration de données adaptées à chaque utilisateur et d'explicitation des processus de traitement (transparence des critères de sélection ou de rejet), nos travaux futurs porteront sur une redéfinition et une exploitation plus intense de méta-données de qualité (indicateurs de fiabilité, pertinence, complétude,...) et sur la mise au point de mécanismes de recommandation adaptatifs aux niveaux de sensibilité/criticité des données et aux points de vue multiples des utilisateurs.

Références

[Belkin, 1992] Belkin N. J. Croft W. B. (1992). Information Filtering and Information Retrieval : two Sides of the Same Coin ? *Communications of the ACM*, **35**(12), 29-38

[Bouzeghoub, 1998] Bouzeghoub M. Gardarin G. Valduriez P. (1998). *Les Objets*, Ed. Eyrolles

[Chen, 1976] Chen P. P. (1976). The Entity-Relationship model : Toward a Unified View of Data, *ACM Trans. Database Systems*, **1**(1), 9-36

[Connolly, 1997] Connolly T. Begg C. Strachan A. (1997). *Database Systems : A Practical Approach to Design, Implementation and Management*, Ed. Addison-Wesley

[Elmasri, 1994] Elmasri R. Navathe S. (1994). *Fundamentals of Database Systems*, 2nd Ed. Benjamin / Cummings

[Graveleau, 1997] Maintenance d'une Base de Données techniques de Référence : l'apport du Veilleur à la fourniture d'information validée aux experts de l'entreprise, *Ile Rousse*, 1997

[Gardarin, 1996] Gardarin G. & Fils (1996). *Le Client-Serveur*, Ed. Eyrolles

[Irwin, 1995] Irwin N. H. DeLand S. M. Crowder S. V. (1995). *Extraction of Information from Unstructured Text*, Tech. Rep. of Sandia National Laboratories, SAND95-2532

[Pazienza, 1997] Pazienza M. T. (Ed.) (1997). *Information Extraction*, Lecture Notes in Artificial Intelligence, (1299)

[Potter, 1989] Potter W.D. Trueblood R. P. Eastman C. M. (1989). Hyper-semantic Data Modeling, *Data & Knowledge Engineering*, (4), 69-90

[Turtle, 1992] Turtle H. R. Croft W. B. (1992). Comparaison of Text Retrieval Models *Computer Journal*, **35**(3), 279-290

[Ullman, 1988] Ullman J. D. (1988). *Principles of Database and Knowledge-Base Systems*, vol.1/2, Ed. Computer Science Press

Annexe

Recherche et Filtrage d'information : deux nécessités complémentaires

Le filtrage d'information est un terme utilisé pour décrire une variété de traitements mis en oeuvre pour délivrer l'information, bien qu'il soit relativement populaire grâce aux services de courrier électronique ou pour les systèmes multimédias distribués, la distinction entre le filtrage d'information et les traitements tels que la recherche (*Information Retrieval* IR) et l'extraction d'information (*Information Extraction* IE) n'est pas très claire [Belkin, 1992]. Nous nous contenterons de souligner leur proximité et leurs divergences et, de constater qu'elles répondent à deux des besoins essentiels en terme de traitement de l'information dans une application de veille technologique.

Le concept de la recherche d'information
Lorsqu'un utilisateur est à la recherche d'une information, il soumet volontairement une requête à un système dans le langage de requêtes particulier du serveur. Le sens des textes est représenté par indexation plus ou moins complexe des termes et la comparaison entre la requête et ces mots-clés ou concepts indexés permet la sélection des textes présumés pertinents. Ces textes trouvés sont ensuite évalués par l'utilisateur. Cette évaluation entraîne souvent une modification de la requête et son raffinement progressif: c'est le processus de *relevance feedback* (réinjection des nouveaux critères de pertinence découverts lors de l'étape précédente).

Le concept du filtrage d'information
Le filtrage d'information s'intéresse à des utilisateurs qui ont des besoins en informations relativement stables sur le long terme et une régularité dans leurs intérêts en termes d'informations (ex. ils ont besoin d'être au courant des évolutions et mises à jour sur un sujet précis) et ne changeront pas ou lentement de critères de recherche.
Ces utilisateurs auront un comportement plus passif vis-à-vis de la recherche d'information.

	Recherche d'information	**Filtrage d'information**
Utilisation	ponctuelle par un utilisateur ayant un but et une requête uniques et ponctuels	répétées par un (ou des utilisateur) ayant des intérêts et buts à long terme
Paradigme	admet les problèmes d'adéquation entre les représentation des besoins en information et les requêtes	basé sur les préférences de l'utilisateur prend l'hypothèse que les profils sont des spécifications correctes des intérêts en information
Orientations	représentation des textes, de leur organisation, de la comparaison textes et termes recherches et modification des requêtes	représentation des intérêts en tant que profils ou requêtes et à la distribution des textes aux individus
Méthodes Objectif	expression de ce qui est recherché dans un langage de requête sélection de textes au sein d'une base relativement statique	expression de ce qui est recherché et de ce qui ne l'est pas sélection ou l'élimination des textes dans un flux dynamique de données
Modèles	□ modèle bouléen basé sur le principe de l'appariement exact des termes recherchés et avec les mots-clés des textes trouvés □ modèle de type espace vectoriel ou modèle probabiliste basés sur le principe d'un appariement optimal avec 1) la pondération des termes recherchés (par la fonction *tf.idf*) ou celle des mots-clés dans les textes indexés 2) l'ordonnancemenrt des textes selon une probabilité de pertinence par rapport à la requête [Turtle, 1992]	modèle de graphes d'inférence ou arbres de décision Bayesien pour le calcul des probabilités sur la comparaison simultanée d'une nouvelle information à plusieurs profils.
Visée	répondre à l'utilisateur par des textes dans un épisode de recherche unique	s'intéresser aux changements long terme sur une série d'épisodes de recherche d'information
Aspects contextuels		
Public ciblé	une communauté d'utiliseurs est bien définie ayant des domaines d'intérêts spécifiques	un utilisateur ou un groupe d'utilisateurs non préalablement définis et ayant des domaines variés d'intérêts
Fraîcheur d'un texte		primordiale
Problèmes soulevés		intrusion ou d'atteintes à la vie privée

Tableau 1

Du point de vue de la veille technologique, ces 2 fonctions s'apparentent à la distinction entre la surveillance de sujets biens délimités dans le but de détecter rapidement les nouveautés (*push*) et la prise de connaissance d'un domaine peu connu (*pull*).