Documents, données et méta-données : une approche mixte pour un système de veille

Laure Berti-Équille, David Graveleau

Documents, données et méta-données : une approche mixte pour un système de veille

Laure BERTI-EQUILLE (*), **David GRAVELEAU** (**)
berti@irisa.fr, david.graveleau@ctsn.dga.defense.gouv.fr

(*) IRISA, Campus Universitaire de Beaulieu, 35042 Rennes cedex, France
(**) CTSN, B.P. 28, 83800 TOULON Naval, France

Mots clefs :
sources d'information, informations textuelles, données structurées, méta-données, analyse et conception d'un système de veille

Keywords:
information sources, textual information, structured data, metadata, watch system modeling and design

Palabras clave :
fuentes de información, información textual, datos estructurados, méta datos, métodos de diseño y diseño de los sistemas de vigilancia

Résumé

L'exploitation de grandes masses documentaires pour l'élaboration d'un dossier de veille technique nécessite la mise en œuvre d'un système d'information adapté à la compilation de données multisources. Le retour d'expérience sur l'utilisation du système de veille SILURE au Centre Technique des Systèmes Navals, système présenté dans nos précédentes contributions [Gra97,BG98], nous conduit à étendre la modélisation initiale pour une meilleure prise en compte du contexte documentaire d'où sont extraites les données sélectionnées. L'originalité de cette double approche (« orientée donnée » et « orientée document ») repose sur l'emploi de méta-données relatives à la qualité des données stockées et à celle de leurs sources (intérêt, fiabilité, complétude, fraîcheur). L'exploitation combinée de ces méta-données permet notamment d'affecter les priorités de traitement sur une collection de documents qui va, par une structuration sélective semi-automatique, assurer l'alimentation en données factuelles et référentielles de la base au cœur du système de gestion des informations du domaine ciblé par la démarche de veille.

Introduction

Pour être un support efficace de la prise de décision stratégique, la note de synthèse accompagnant tout dossier d'information remis à la Direction de l'entreprise doit présenter l'essentiel des faits établis et des enjeux identifiés par l'équipe en charge de la veille sectorielle. A l'opposé, l'élaboration du dossier d'information proprement dit est, par nature, le produit composite de multiples opérations de collecte de publications, de comptes-rendus, de missions ou d'avis d'expert, puis de sélection et d'ordonnancement des documents que l'on juge constituer l'image la plus fidèle de l'environnement concurrentiel et technologique.

Quand le consensus tarde à se faire autour d'une décision voire d'une orientation stratégique, il n'est pas rare de voir mis en cause quelques-uns des éléments présentés dans le dossier de veille au motif qu'ils ne reflètent pas correctement ou complètement la réalité. Il est alors nécessaire de revenir à la source même de la donnée contestée voire de reprendre la recherche en en élargissant le cadre à la marge du secteur étudié. La traçabilité de l'information, qui consiste à conserver en mémoire le contexte de production et de diffusion de toute donnée jugée pertinente par le veilleur, constitue le socle de sa démarche de fusion progressive de l'information multi-source destinée à la note de synthèse : à partir de données de valeur et d'intérêt inégaux du fait de leur caractère souvent lacunaire, incertain et surtout contradictoire, le veilleur opère sélections, analyses critiques et validations sur les données brutes disponibles afin d'en tirer l'information élaborée qu'il souhaite présenter. Quand la masse d'informations est importante et que le travail de synthèse se fait en équipes de spécialistes sectoriels, il est nécessaire de disposer d'un système d'information capable de constituer le référentiel de données, de connaissances et de documents accumulés sur les différents dossiers thématiques couverts par les veilleurs.

Centre technique en charge de la spécification et de l'évaluation des matériels destinés à la Marine Nationale, le CTSN (Centre Technique des Systèmes Navals) a conçu et met en œuvre le logiciel SILURE pour la constitution et l'entretien du référentiel technico-opérationnel dans les domaines se situant au cœur de ses métiers d'étude et d'expertise des systèmes navals.

Outil de dialogue entre le veilleur et l'expert [Gra97], il vise, d'une part à mémoriser tous les constituants des dossiers d'informations dont les veilleurs ont la charge et à simplifier la production des notes de synthèse sur les technologies, produits et acteurs navals. Composé d'une base de données structurées et de modules de criblage d'information, ce système de veille mémorise à la fois 1) les fiches descriptives des plates-formes, des équipements navals et des acteurs (concepteurs ou utilisateurs de ces systèmes), 2) les tableaux de synthèse qui les mettent en relation ou les comparent et 3) les documents d'où sont tirées ces données. Avec la multiplication des éditeurs d'information électronique et l'augmentation continue des flux de données qu'ils diffusent, les procédures successives d'extraction, de structuration et de mise en base de l'information factuelle (approche « orientée donnée ») tendent à générer des files d'attente et limitent la réactivité du système de veille. Par ailleurs, son ouverture aux contributions de tous les acteurs de l'entreprise impose de développer son aspect collaboratif, notamment en rendant plus transparente l'interprétation indissociablement liée à la sélection et la recommandation d'une donnée.

Dans une perspective de rééquilibrage des efforts investis pour la validation des données factuelles au profit de l'intégration immédiate et de la mise à disposition des informations textuelles de nature événementielle ou référentielle recueillies en continu [BG98], nous envisageons par complémentarité l'approche « orientée document » en développant les mécanismes de classification multicritère sur la qualité des documents (fraîcheur, fiabilité, intérêt métier, couverture et profondeur descriptive) afin d'établir leur niveau de priorité de traitement et d'entrée dans la base. Pendant des critères de qualification de la donnée (cotation), les méta-données de qualité viennent progressivement enrichir le document-source et permettent de conserver les avantages liés à l'approche « orientée donnée » tout en garantissant le maintien de la richesse contextuelle propre à l'approche documentaire.

1 Retour d'expérience du système de veille SILURE

1.1 Définition du besoin

La spécification des objectifs de performance des futurs systèmes navals de la Marine Nationale est un processus itératif qui tente de concilier le besoin des utilisateurs et l'état de l'art technique dans chacun des sous-domaines fonctionnels concernés. *In fine*, une approche proprement 'système' est mise en œuvre pour assurer la cohérence des choix unitaires dans le respect des grands équilibres du projet (efficacité globale, coûts, délais).

Dans un contexte de forte pression budgétaire et d'accélération de la construction de la Défense européenne, il est crucial d'être en mesure d'évaluer très en amont du projet les enjeux que représentent l'emploi ou non de telle ou telle technologie émergente et le développement de tel ou tel concept innovant. Le Plan Prospectif à 30 ans (PP30) de la DGA (Délégation Générale aux Armements) et le Plan Stratégique de la DCE (Direction des Centres d'Expertise et d'Essai) qui s'en inspire entendent définir les capacités à développer tant vis-à-vis de la menace que de l'environnement de coopération ou de concurrence européen. Ces outils d'orientation stratégique requièrent, pour leur entretien permanent, que soit mise en place une démarche de surveillance et d'analyse de l'environnement technologique et géostratégique mondial dans tous les secteurs d'activité concernant la Défense. Pour les systèmes navals, le CTSN a reçu la mission de constituer le référentiel des connaissances sur les acteurs et systèmes étrangers, nécessaire à l'élaboration des documents d'analyse prospective ou de définition des futurs systèmes. Le système SILURE a ainsi été conçu dans un objectif de certification des données technico-opérationnelles à la base des grands choix techniques pour les programmes de sonars : s'il ne prétend pas répondre aux questions d'ordre stratégique (par exemple : quel niveau de performance atteindre pour faire face à telle menace ? De combien de frégates la Marine doit-elle être équipée ?), les données de référence qu'il fournit sont cependant celles utilisées dans la plupart des études technico-fonctionnelles qui jalonnent le processus de définition des équipements navals futurs. Comme tout bon référentiel de veille, il permet l'évaluation rapide des potentiels actuels et donne une tendance pour leur développement prévisible : de quoi sont et seront dotés les partenaires (interopérabilité des alliés) et les adversaires potentiels (maîtrise des risques) ? Que proposent sur le marché les principaux industriels (chantiers navals et équipementiers) ? Quelles technologies sont employées dans les grands programmes concurrents ?

1.2 Choix de la modélisation des données

Pour constituer sur le long terme un tel réceptacle de connaissances multi-domaines, la conception du système de veille SILURE s'est, dès l'origine, posé le problème de la modélisation des données à mémoriser [BG98]. Deux grandes directions s'offraient traditionnellement[1] (Tableau 1):

- l'approche « orientée donnée » privilégiant l'information factuelle, avec une base de données structurées selon les entités du métier et les relations qu'elles entretiennent,
- l'approche « orientée document », privilégiant l'information référentielle avec la constitution d'une banque de documents indexés.

Type d'Approche ou	**Objet Central (Focus)**	
Niveau d'information	**Entités à modéliser**	**Données d'entrée**
Factuel	Systèmes navals & Acteurs	Données descriptives des entités
Evénementiel	*Historique des relations entre Systèmes & Acteurs*	*Récits d'actualité mettant en scène les Entités ciblées par la veille*
Référentiel	Documents Sources & Emetteurs	Références bibliographiques

Tableau 1. Cibles d'information selon l'approche de modélisation

[1] Dans notre précédente publication [BG98], nous envisagions aussi l'approche « orientée actualité » ou événementielle, correspondant à la mémorisation de l'historique des relations entre les entités objet de la veille (récits d'actualité mettant en scène les acteurs et systèmes du domaine). Il s'avère qu'aucun système informatique ne propose actuellement la formalisation de ce type d'approche. Ce sera une direction de recherche pour nos travaux futurs.

1.3 Approche «orientée donnée» versus approche «orientée document»

Chacune de ces approches classiques possède ses avantages et ses inconvénients.

- **Avantages des approches**
  - Pour l'approche orientée donnée, on soulignera :
    - sa proximité par rapport au besoin final de synthèse, le travail de sélection et de présentation de toutes les données requises ayant été fait en amont par l'analyste.
    - sa clarté dans la représentation des entités et de leurs multiples relations (réseaux de liens typés et arborescences d'entités), un grand effort de formalisation et de validation ayant été fourni pour désigner précisément les entités décrites et renseigner correctement les paramètres via les attributs de la base.
    - sa capacité à comparer rapidement les entités entre-elles ainsi que la richesse et la complexité des requêtes possibles.
  - Pour l'approche orientée document, on soulignera :
    - sa plasticité formelle (respect de la variété de formats et de structurations sous laquelle une information est présentée par sa source).
    - sa richesse d'évocation (stockage d'informations signalétiques, descriptives sur les émetteurs et le contenu des documents, événementielles sur la vie des acteurs et des systèmes).
    - sa rapidité d'intégration d'une information nouvelle (ne nécessitant que peu de traitements en amont).
- **Inconvénients des approches**
  - Pour la base de données, on mettra l'accent sur :
    - la nécessité de définir *ab nihilo* le modèle conceptuel de données et la difficulté à faire évoluer le schéma de la base correspondant pour se conformer aux nouveaux aspects des entités à décrire
    - son incapacité à conserver sans perte le contexte de production et de diffusion de l'information (rupture du lien vers la source de la donnée).
    - sa difficulté à conserver une information parcellaire (stockage de valeurs incomplètes), approximative (représentation de valeurs imprécises) ou contradictoire (cotation de l'incertitude et des incohérences).
  - Pour la base documentaire, on mettra l'accent sur :
    - son incapacité à délivrer rapidement un état correct et à jour sur une question, la dimension historique (fraîcheur) et parcellaire (inexistence d'un catalogue complet de toutes les données requises) des documents venant brouiller la vision globale du veilleur.
    - sa difficulté à stocker des informations structurées selon les entités ciblées.

En résumé, ces deux types de systèmes d'information possèdent les avantages de leurs inconvénients (Tableau 2) : la formalisation excessive de la base de données peinant à préserver l'orientation intrinsèque et la richesse expressive des documents originaux, cette dernière étant, quant à elle, un obstacle à la synthèse rapide des traits caractéristiques des entités à décrire et à comparer.

Type d'Approche ou **Niveau d'information**	**Système d'Information**	
	Type	**Points forts des représentations**
Factuel	Base de Données	Richesse du modèle Entités-Relations
Evénementiel	*Prototypes de bases de données temporelles*	*Sémantique & Historique des relations entre entités*
Référentiel	Base Documentaire	Complexité Thématique & Préservation de l'expression originale

Tableau 2. Type de système d'information selon l'approche de modélisation et de gestion des données

1.4 Le modèle de conception de SILURE

Afin de conserver le meilleur des deux mondes à la fois sur les plans conceptuel et physique, il a donc été décidé de concevoir SILURE autour [GRA97] :

1. d'un SGBD (relationnel étendu) mettant en oeuvre un méta-modèle des données acceptant des attributs (multi)valués et des valeurs référencées et cotées :
   - le méta-modèle garantissait l'évolutivité du modèle métier, en permettant la description progressive des entités par assemblage d'attributs prédéfinis,
   - les attributs décrivant une entité ou une relation entre entités acceptaient par nature un nombre in(dé)fini de valeurs simples ou complexes,
   - les valeurs dites "brutes" (VB) étaient référencées (lien direct vers l'entité Source de la donnée) et cotées (affectation d'une note de qualité définissant la vraisemblance de la valeur). Après criblage à l'aide des outils décrits ci-après, l'utilisateur pouvait finalement sélectionner une valeur dite "recommandée" (VR), réputée la "meilleure" donnée disponible.
2. d'un ensemble de modules de criblage de l'information destinés à tester la validité des données et renforcer la cohérence et la complétude de la base :
   - le module de détection d'incohérences entre :
     - les VB concurrentes pour un même paramètre (valeur d'un attribut d'une entité)
     - les VR pour des paramètres liés (relation logique, physique ou empirique)
     - les cotations concurrentes pour un même paramètre ou des paramètres liés
   - le module d'émission d'hypothèses pour les VR manquantes (restitution) et de proposition de leur cotation
   - le module d'inspection de l'état de renseignement de la base (statistiques croisées entre les entités, les attributs, les valeurs VB / VR et les références).
   - Le module de recommandation de VR selon les cotations, en tenant compte de la fraîcheur des VB associées et des niveaux de protection des données.

Ce faisant, la notion de méta-données (ou 'données sur les données') a été introduite en distinguant : les méta-données de qualité (Cotation) ou celles indirectement liées à l'évaluation de celle-ci :

- le référencement de chaque valeur autorisait la traçabilité des VR, en permettant à tout instant de remonter la chaîne de production et de diffusion des VB,
- il permettait aussi, à l'inverse de juger de la complétude d'une référence (en termes de nombre de VB fournies, que ce soit en profondeur descriptive ou en couverture thématique).
- la cotation, enfin, était étendue à tous les niveaux de la base, comme indicateur de fiabilité : cotation des VB et des VR, des références et de leur émetteur, autorisant de cette manière des opérations de contrôle de cohérence ou d'agrégation de cotation aux niveaux supérieurs.

Cette modélisation très riche par son expressivité et son caractère générique permettaient ainsi de stocker conjointement toutes les données constituant l'ensemble des dossiers de veille, qu'elles soient issues de documents en entrée ou élaborées par l'équipe de veille. Elle autorisait la fourniture à tout instant d'une valeur recommandée (VR) cotée et susceptible d'être accompagnée d'un dossier justificatif référençant : sa source, les VB concurrentes issues des différentes sources disponibles et des traitement liés au contrôle de la cohérence ou à l'application de règles de vraisemblance.

Ces objectifs ont partiellement été atteints, deux obstacles s'opposant néanmoins à leur réalisation :

- l'imperfection du modèle de données, qui ne permettait pas de transcrire assez fidèlement l'information brute (valeurs approchées, lien historique entre les VB et les VR) et d'exprimer assez finement la qualité des données sélectionnées (VB) puis recommandées (VR),
- l'inadéquation et le vieillissement des outils informatiques de la plate-forme d'origine (SGBD Ingres, Logique fonctionnelle et IHM sous Lisp) qui, liés à l'accroissement permanent du volume des données stockées, allongeaient les temps de réponse de la base et ne permettaient plus son entretien normal. Qui plus est, il devenait impossible de résister à l'accélération continue du nombre de publications à intégrer.

2 Document, données et méta-données : quelle nouvelle modélisation pour SILURE ?

A l'occasion de la refonte de la plate-forme informatique SILURE, il a donc été envisagé une mise à plat complète du système, chacun de ses aspects et de ses composants étant jugé à l'aune des services qu'il rendait. Outre ce bilan fonctionnel, une étude bibliographique sur les méta-données liées à la qualité de l'information a permis de dresser un panorama de ce champ de recherches et d'investigations et d'étudier son application potentielle à notre système de veille.

2.1 Spécification et exploitation des méta-données : un état de l'art

De plus en plus employées pour la recherche d'information dans les documents multimédias, le texte et les bases de données structurées (notamment, comme aide à l'interopérabilité), les méta-données ont été successivement définies et employées dans des contextes fort différents :

- Chen et al. [CHK+94] ont défini les méta-données comme des propriétés dérivées des documents ou des médias, utiles pour l'accès ou la recherche de l'information.
- Dans le projet Meta-database de [Hsu91], une approche de gestion des méta-données est adoptée pour réaliser une synergie globale entre différentes bases de données composant une architecture distribuée.
- Kiyoki et al. [KKH94] mettent en application une recherche d'images par association sémantique basée sur les méta-données liées aux mots-clés qui représentent l'impression de l'utilisateur face au contenu des images. De même, Anderson et Stonebraker [AS94] ont développé un schéma de méta-données pour des images satellites ; Jain et Hampapur [JH94] ont proposé des méta-données constituant une représentation intermédiaire de l'information audiovisuelle.

Première tentative de classification intéressante, Bohm et Rakow [BR94] ont établi un recensement et une typologie des méta-données selon leur nature et les différents buts recherchés :

- *les méta-données basées sur la représentation des types d'information* : dans [AS94], les méta-données sont employées pour supporter différentes perspectives pour l'interprétation des images satellites : selon le point de vue de l'informaticien ou selon celui du météorologue.
- *les méta-données basées sur le contenu* : méta-données extraites à partir de l'information physique de la donnée, dont on trouve un exemple dans les vecteurs associés aux documents textuels au sein d'une collection de documents indexés selon une méthode d'indexation particulière.
- *les méta-données descriptives de contenu* : elles ont été utilisées dans [KS94] pour la découverte et la recherche d'informations dans des multibases. Les dispositifs dépendants de domaine et les méta-dispositifs indépendants du contenu dans [JH94] sont des exemples de méta-données descriptives de contenu pour l'information audiovisuelle.
- *les méta-données basées sur la composition ou la localisation du document* (et créées au moment de l'extraction),
- *les méta-données basées sur les conditions de présentation des données ou des documents* (par exemple, l'algorithme d'ordonnancement selon un pourcentage de pertinence peut constitué une méta-donnée, tout comme la présence d'une information dans un titre)
- enfin, *les méta-données basées sur la collection des documents et la navigation possible entre documents ou items d'informations*.

2.2 Méta-données et qualité de données

L'utilisation de méta-données pour l'amélioration de la qualité des données, quant à elle, a été préconisée dans [Rot96] où les producteurs d'informations sont incités à assurer la vérification, la validation et la certification de leurs données (VV&C : *Verification, Validation and Certification*). Les méta-données concernant la qualité des données devant être fournies de pair avec les données pour permettre l'estimation et la maintenance de la qualité des données. Dans le domaine des Systèmes d'Information Géographique, la plupart des standards d'échange (ISO, CEN, EDIGéO, FGDC), incluent les spécifications de méta-données de qualité [GJ98]. Les plus anciens travaux sur la

modélisation de la qualité des données ont été menés par [Bro80]. Dans [BP85], les auteurs ont, les premiers, défini la notion de qualité des données et ses dimensions. La terminologie et les concepts fondamentaux sur ce thème ont également été précisés dans [FLR94]. De nombreuses propositions ont depuis été faites [TB98] [SK97] [CP98] et le consensus sur la définition de la qualité des données n'est toujours pas atteint [WSF95]. Certains travaux intègrent totalement la modélisation et la gestion de la qualité des données dès la conception d'un système d'information. Les premières approches adoptées pour mesurer la qualité des données furent des approches statisticiennes centrées sur des méthodes telles que l'inférence sur les données manquantes à partir des modèles statistiques, la détection automatique et le traitement des exceptions et des données isolées [LU90]. De nombreuses méthodes ont été développées pour mesurer la qualité des données fournies aux utilisateurs conformément à leurs propres spécifications de qualité.

En synthèse, les différents travaux sur la qualité des données peuvent être classés en trois grands courants selon leur objectif qui est : 1) soit de définir rigoureusement chaque dimension de la qualité des données ainsi que leur mode de mesure sur la base d'une méthode scientifique, 2) soit de créer un ensemble standard universel de dimensions opérationnelles pour la qualité des données, 3) soit de proposer une assistance à l'utilisateur pour qu'il définisse et évalue lui-même la qualité des données qu'il manipule.

Pour la qualification des données et des documents gérés par SILURE, on a dressé une liste de critères de qualité et d'indices liés, à partir des principales mesures de qualité de données proposées dans la littérature [Red96] [FLR94] [RW95] [SK97]. Au nombre de quatre, chacun de ces critères est valué par la mesure de la valeur d'un attribut, d'une facette d'attribut ou d'une relation entre attributs :

- la **fiabilité** combine **:**
  - une estimation de l'**exactitude,** qui se mesure en détectant le taux de valeurs incorrectes dans la base de données (en fait inférieur à un seuil de cotation pour les VR)
  - un test de **cohérence** par rapport à un ensemble de contraintes en détectant les données de la base qui ne les satisfont pas (tests par des règles purement logiques, le respect de lois physiques ou des 'lois' empiriques issues de la détection de régularités).
- la **complétude** se mesure en détectant le taux de valeurs manquantes dans la base de données (taux de remplissage par comparaison du nombre de valeurs canonique pour l'ensemble des entités avec le nombre effectif)
- la **fraîcheur** se mesure en détectant d'une part, le taux de valeurs "obsolètes" dans la base de données, et d'autre part, l'âge moyen des descriptions d'entités (par observation des dates de références),
- l'**intérêt**, enfin, qui se mesure en comparant le taux de paramètres dits 'importants' (ciblés par les recherches) renseignés sur les non renseignés.

2.3 L'extension du modèle SILURE

Le bilan fonctionnel de SILURE a, quant à lui, conduit à conforter l'approche factuelle d'origine, tout en suggérant des extensions vers l'approche documentaire. L'utilisation poussée des fonctions de comparaison d'entités grâce à la mise en relation des valeurs d'un même attribut était renforcée par l'immensité du champ ouvert par la multivaluation référencée : à chaque niveau de la modélisation, des statistiques spécifiques viennent renforcer et la complétude et la cohérence des données compilées (nombre et cotation moyenne des VB d'un attribut, des VR d'attributs logiquement dépendants, des VR des attributs d'une même entité ou d'entités décrits par la même référence,…). Pour prolonger encore les efforts de formalisation de l'information et améliorer l'expressivité de la représentation ainsi définie, on a cependant procédé à l'enrichissement du méta-modèle, sous trois aspects :

- la mise en œuvre et l'utilisation d'un SGBD orienté objet a permis de développer la description arborescente du modèle métier et, en faisant concilier l'approche utilisateur (vision distincte 'entité' et 'relations entre entités') et la représentation informatique interne, a apporté des gains de performance considérables. Elle a permis de même de bénéficier véritablement des avantages du

méta-modèle en autorisant l'administrateur de données à remodeler en permanence le modèle métier (afin de s'adapter aux nouvelles données) sans toucher au schéma de la base.

- la distinction entre cotation d'une Valeur Recommandée (de "sûr" à "faux", en passant par "probable", "possible", "douteux") et la notation d'une Valeur Brute (sur une échelle étendue de -1 à 20) établit une distinction entre la précision interne sur la fiabilité de la donnée et le niveau de confiance publié quand celle-ci est diffusée par l'équipe de veille.
- l'ajout d'une zone de commentaire pour chaque valeur, pis-aller permettant de ne pas étendre à l'infini la liste des critères de qualité à indicer pour chaque valeur (lourdeur de saisie), tout en permettant d'insister sur le trait qui semble le plus important : l'intérêt dû à la fraîcheur ("nouveau !") ou l'originalité d'une donnée ("rare !"), son imprécision relative ("valeur approximative", "de l'ordre de…"), le caractère incomplet, incertain ou incohérent explicitant sa faible notation, ou la nature même de la donnée (mesure, hypothèse, spécification à atteindre,…).

NOUVELLE MODELISATION DU SYSTEME DE VEILLE SILURE

Schéma physique	**Méta-modèle**	**Modèle métier**	***Exemple***
Classe	Catégorie	Arborescence de Catégories métier	*Plate-forme navale > Sous-Marin*
Objet	Entité	Objet métier	*Kursk*
	Référence	Document	*Le Monde*
Champ	Attribut simple	Paramètre	*Longueur*
	Attribut structure	Relation	*Endurance (distance-temps)*
Valeur	Valeur (multivaluée)	Valeur Brute	*L:120m [12] Libération* *L :118m [8] Le Monde*
		Valeur Recommandée	*L :120m [B]*
	Méta-données	Cotation / Notation	*Echelles : de A à F / de 0 à 20*
		Lien sur Référence	*Le Monde (31/12/1999)*
		Protection	*Diffusion Restreinte*
		Commentaire	*Valeur rare !*
		Lien VR / VB	

Tableau 3. Modélisation des données retenue dans la dernière version de SILURE

Cette dernière modification a conduit *de facto* à un rapprochement avec l'approche documentaire, en enrichissant la simple transcription de la valeur brute de la donnée par une appréciation plus fine fournie par l'analyste l'ayant sélectionnée. Cependant, l'approche formaliste de mise en base de données structurées étant conservée intégralement, il reste à mieux restituer le contexte de production de la valeur et faciliter l'intégration progressive des documents dans le système d'information.

3 Du document à la donnée (et inversement) : les avantages d'une approche mixte

Une modélisation associant données et documents via les méta-données ayant été définie, il reste à spécifier les outils pour son exploitation :

3.1 Vers un système intégré

Idéalement, on perçoit un système dual qui, à partir du document original, laisserait le choix entre :

- une représentation naturelle progressivement enrichie de surcouches de méta-données de qualité,
- une représentation structurée des données référencées telle que celle que l'on constitue actuellement par la mise en base des données sélectionnées et recommandées.

Le passage entre ces deux représentations conduirait à la mise au point de deux outils :

- pour la saisie, un module de **structuration sélective par surlignage et d'annotation** permettant, à partir des entités et des attributs détectés dans le texte original, d'affecter chaque valeur à un attribut d'une entité puis de lui affecter des méta-données de qualité.

- pour la consultation, un module qui présente alternativement :
  - une vision **naturelle**, plus ou moins surchargée des attributs et entités et d'autres méta-données de qualité,
  - une vision **synthétique** correspondant à la représentation structurée telle que présentée par la base SILURE actuelle :
    - fiche d'entité (ensemble d'attributs)
    - tableau ou graphes de relations entre entités
    - tableaux de comparaison entre attributs d'entités,

  Ces 3 représentations étant, elles aussi, plus ou moins complétées par des méta-données de qualité initialement définies (VR, VB, cotation/notation,…).

3.2 Le retour aux sources : une solution intermédiaire

Face à la difficulté de définition et de réalisation des modules de surlignage et de vision documentaire multicouches, on évaluera au préalable une solution intermédiaire, consistant à ajouter une étape dans la navigation actuelle entre la donnée et sa source (actuellement réduite à sa simple description), celle de la navigation préalable vers le paragraphe d'où est issue la donnée. On donnera ainsi à l'utilisateur un 4ème niveau progressifs d'approfondissement de son questionnement sur la valeur de la donnée qui lui est recommandée :

- dans quelle mesure la valeur a-t-elle été recommandée : quelle est sa cotation en tant que meilleure valeur ou son classement parmi toutes celles qui me sont accessibles, qui l'a recommandée ?
- de quelle valeur brute est-elle issue : quelle est sa notation initiale, sa fraîcheur, sa source (qui l'a noté, quelle est sa cotation, quelle bibliographie) ?
- quels sont les traitements de validation ou de restitution qui lui ont été appliqués ?
- dans quel contexte précis est-elle citée (visualisation de la formulation exacte dans le paragraphe considéré) ?

3.3 Un ordonnanceur pour le traitement des documents

La première réalisation spécifiée est un module destiné à affecter un ordre de priorité pour la saisie des documents et leur structuration progressive.

Dans un premier temps, des méta-données renseignant des critères tels que la fraîcheur, la fiabilité, l'intérêt, la couverture et la profondeur descriptives des documents sont définies de façon semi-automatique (en partie par les veilleurs, experts et documentalistes).

A titre d'exemple, on définit ci-après une typologie de documents selon leur (Tableau 4):

- richesse descriptive : critères de quantité d'Objets et de quantité de paramètres et Valeurs à saisir
- structuration formelle : critères de degré de Formatage.

Cette grille va permettre d'attribuer des pondérations à chaque type de document à traiter :

	Critère / Niveau	Elevé (1)	Moyen (2)	Faible (3)
O	Quantité d'objets	>50	5-50	<5
V	Quantité de paramètres	>30 / objet	3-30 / objet	<3 / objet
F	Degré de formatage	Fiche	Texte structuré	Texte libre

Tableau 4. Typologie des documents sur des critères quantitatifs et formels

Il en découle une classification a priori, avec les équivalences associées (Tableau 5) :

Type	Document à analyser (et volume typique)	Caractéristiques O - V - F	Nb unités (source papier)	Nb unités (source électronique)
A	Article (qq pages)	03 / V2 / F3	2	1
B	Notice 'produit' (qq pages)	03 / V1/ F2	3	2
C	Catalogue 'produit' (qq dizaines de pages)	02 / V2 / F1	5	3
D	Base de données / recueil de fiches (qq centaines de pages)	01 / V1 / F1	3	2
E	Listing (1-4 pages)	01 / V3 / F1	2	1

Tableau 5. Classification des documents selon leur nature, volume, type et support

Les méta-données générées automatiquement ou renseignées de façon manuelle vont permettre de qualifier les documents et de réordonner leur traitement, non plus selon leur ordre d'arrivage et de prise de connaissance par les veilleurs mais selon un ordre de priorité qualitative (Figure 1). Les documents seront structurés de façon sélective d'après ce nouvel ordre afin de mettre en valeur l'information utile (ou information d'intérêt) et la structurer à destination de la base de données.

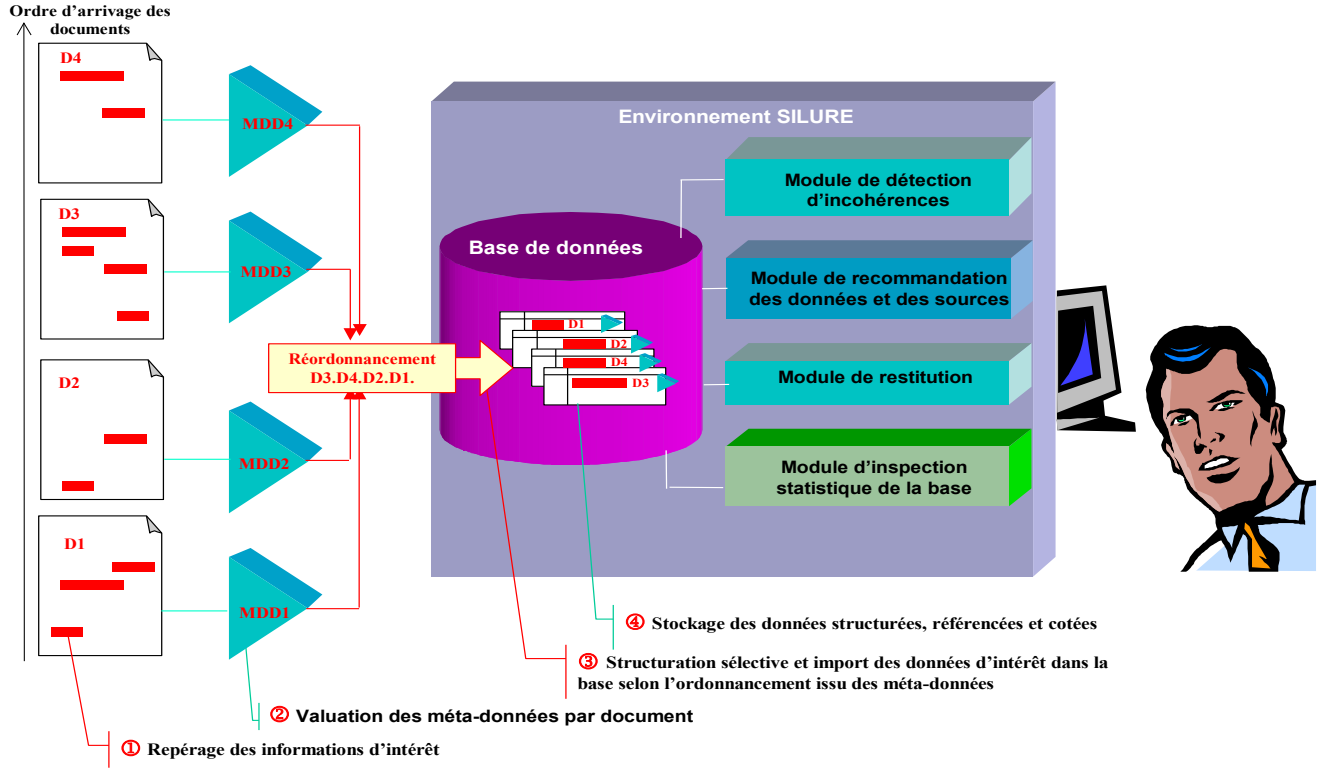

Figure 1. Chaîne de traitement mixte : des documents vers les données et inversement

Les données extraites et stockées hériteront à la fois des considérations qualitatives relatives à leur source référencée et d'une expertise qualifiant chaque donnée selon une cotation (indicateur de vraisemblance) accompagnée d'un dossier justificatif. Les différents modules d'analyse (restitution, détection des incohérences, inspection statistiques) décrits précédemment veilleront au criblage et à l'épuration des données de la base. A l'autre extrémité de la chaîne de traitement, lors de la consultation du système de veille, une recommandation des données a priori les plus vraisemblables et/ou des documents jugés les plus fiables à la date de la recherche sera proposée à l'utilisateur.

4 Conclusion

Système d'information destiné à faciliter la veille collaborative dans les multiples domaines liés à la conception des systèmes navals, SILURE se présente comme un système d'information intégré capable de compiler au sein d'un même référentiel, d'une part l'information élaborée issue du processus de validation, de restitution et de synthèse des données, et d'autre part l'information brute telle que présentée par sa source documentaire. Des méta-données viennent progressivement qualifier l'intérêt, la fraîcheur, la complétude et la fiabilité supposés de chaque valeur pertinente retenue par l'analyste. Un bilan fonctionnel et un état de l'art sur les méta-données de qualité ont conduit à remettre en cause la modélisation de l'information du système d'origine, sans cependant en modifier profondément l'orientation primitive. Mettant en avant l'extrême richesse potentielle d'une base de données structurées multisources et cotées, notamment en terme d'évaluation et de contrôle de la cohérence et de la complétude des informations stockées, il a finalement été décidé une simple extension du méta-modèle original. Si l'ajout d'une zone libre de commentaire associée à chaque valeur s'apparente à un simple fourre-tout limitant au strict minimum la dérive inflationniste des méta-données de qualité, la restauration du lien entre les Valeurs Brutes (VB) et celles finalement recommandées par l'expert (VR) et l'extension du principe de qualification à tous les niveaux du

modèle autorisent désormais la spécification d'un ordonnanceur semi-automatique pour gérer prioritairement le traitement d'un arrivage de documents en attente de saisie. Cette fonction nouvelle entend répondre au défi posé par la multiplication des publications et des sources d'information, notamment électroniques, et pose les premiers jalons vers un véritable système mixte conciliant l'approche « orientée donnée » et celle « orientée document ». Celles-ci se révèlent très complémentaires, selon leur contexte d'exploitation. Il apparaît donc nécessaire de les proposer conjointement à l'utilisateur qui, selon qu'il a besoin d'obtenir rapidement une vue globale d'un ensemble d'entités et des relations qui les unissent (élaboration d'une note de synthèse) ou qu'il souhaite approfondir un aspect particulier du problème qui lui est posé (confection du dossier d'information), pourra alternativement bénéficier de leurs avantages respectifs. Ces dernières spécifications (module de structuration sélective par surlignage et module ordonnanceur) doivent être implémentées pour étendre le système de veille SILURE par toute la richesse de la double approche « orientée document » et « orientée donnée ». On considèrera ensuite l'intégration de la troisième dimension, celle de l'approche événementielle ou « orientée actualité » qui peine encore à trouver sa formalisation intrinsèque.

5 Références

[AS94] Anderson J., Stonebraker M., SEQUOIA 2000 Metadata schema for Satellite Images, SIGMOD Record, special issue on Metadata for Digital Media, December 1994

[BR94] K. Bohm, T. Rakow, Metadata for Multimedia Documents, SIGMOD Record, Dec. 1994

[BG98] Berti L., Graveleau D., Contribution à la définition d'un vigiciel : quelle modélisation de l'information factuelle, événementielle et référentielle ?, Actes du colloque Veille Stratégique, Scientifique et Technologique (VSST '98), octobre 1998

[BP85] Ballou D., Pazer H., Modeling data and process quality multi-input, multi-output information systems, Management Science, vol. 31, no. 2, p. 150-162, 1985

[Bro90] Brodie M. L., Data quality in information systems, Information and Management, vol. 3, p. 245-258, 1980

[CHK+94] Chen F., Hearst M., Kupiec J., Pederson J. , Wilcox L., Metadata for Mixed-Media Access, SIGMOD Record, special issue on Metadata for Digital Media, December 1994

[CP98] Chengalur-Smith I., Pipino L. (Ed.), Proc. of the 3rd Conf. on Information Quality, 1998

[GFS94] Grosky W., Fotouhi F., Sethi I., Content-Based Hypermedia - Intelligent Browsing of Structured Media Objects, SIGMOD Record, Dec. 1994

[GJ98] Goodchild M., Jeansoulin R. (Ed.), Data quality in geographic information : from error to uncertainty, Hermès, 1998

[Gra97] Graveleau D., Maintenance d'une Base de Données techniques de Référence : l'apport du Veilleur à la fourniture d'information validée aux experts de l'entreprise, Colloque Veille, Ile Rousse, 1997

[FLR94] Fox C., Levitin A., Redman T., The notion of data and its quality dimensions, Information Processing and Management, 30(1), 1994

[Hsu91] Hsu C., The Meta-database Project at Renesselaer, SIGMOD Record, 20(4), 1991

[JH94] Jain R., Hampapur A., Representations for Video Databases, SIGMOD Record, Dec. 1994

[KKH94] Kiyoki Y., Kitagawa T., Hayama T., A meta-database System for Semantic Image Search by a Mathematical Model of Meaning, SIGMOD Record, Dec. 1994

[KS94] Kashyap V., Sheth A., Semantics-based Information Brokering, Proceedings of the 3rd International Conference on Information and Knowledge Management (CIKM), 1994

[LU90] Liepins G., Uppuluri V., Data quality control : theory and pragmatics, M. Dekker, 1990

[Red96] Redman T., Data quality for the information age, Artech House Publishers, 1996

[Rot96] Rothenberg J., Les méta-données to support data quality and longevity, Proc. of the 1st IEEE Metadata Conf., 1996

[RW95] Reddy M. P., Wang R., Estimating data accuracy in a federated database environment, Proc. of the 9th Intl. Conf. CISMOD, p. 115-134, 1995

[SK97] Strong D., Kahn B. (Ed.), Proc. of the 2rd Conf. on Information Quality, MIT, 1997

[SLW97] Strong D., Lee Y., Wang R., Data quality in context, Comm. of ACM, 40(5),103-110, 1997

[TB98] Tayi G. K., Ballou D. P., Examining Data Quality, Comm. of ACM, 41(2), 54-57, 1998

[WSF95] Wang R., Storey V., Firth C., A framework for analysis of data quality research, IEEE Transactions on Knowledge and Data Engineering, 7(4), 670-677, 1995