

Management des connaissances augmenté : usage d'un classifieur sémantique pour l'aide à l'élaboration et au maintien en cohérence d'une base de connaissance

François Vexler*, Corentin Mary*, Alain Berger*, Jean-Pierre Cotton*

*Ardans SAS,
6 rue Jean Pierre Timbaud, «Le Campus» Bâtiment B1,
78180 Montigny-le-Bretonneux, France
{fvexler, cmary, jpcotton, aberger }@ardans.fr,
www.ardans.fr

Résumé. L'alimentation en continu des bases de connaissance par les contributeurs est un point essentiel de l'intérêt et de la pérennité de ces dispositifs au sein des organisations. Au-delà de l'effort de production d'un contenu, vient la nécessité de son intégration au sein d'un existant en garantissant le maintien de la cohérence du dispositif. Ce dernier peut s'avérer complexe, d'autant plus qu'il est souvent à vocation pluridisciplinaire.

Les bases étant construites à l'aide de l'outil Ardans Knowledge Maker® (AKM), et donc fortement structurées en termes de modèles et d'arborescences de classification (ou vues), la méthode habituelle s'appuie sur l'expertise du contributeur. Ce mode de fonctionnement peut toutefois être mis en défaut notamment lors de contributions occasionnelles ou encore lorsque l'on souhaite intégrer des contenus de manière automatisée.

Se ramenant à un problème de classification automatique, Ardans utilise la sémantique vectorielle pour mettre en œuvre un processus de classification guidé par les connaissances. Cette approche constitue une aide appréciable au maintien en cohérence.

Après un rappel des principales techniques du domaine, cet article discute de la réalisation faite et de son utilisation dans un récent «Proof Of Concept» pour la classification automatique de nouvelles entrées de retour d'expérience (REx). Ces travaux sont issus de l'industrie et de la vingtaine d'années d'opérations d'ingénierie des connaissances réalisée par Ardans en France et en Europe. Ils préfigurent l'usage prochain du module compagnon d'AKM appelé «Semantic Analysis».

1 Introduction

L'ingénierie des connaissances propose aux organisations de s'appuyer sur une méthode et des outils afin de modéliser et structurer leurs métiers et les connaissances de leurs experts. Depuis sa création, Ardans conforte sa méthode Ardans MAKE® [Berger et al. (2009)] au

fil de ses opérations industrielles et implante dans sa plate-forme Ardans Knowledge Maker® (Ci-après AKM) [Besson et Berger (2015)] les fonctionnalités qui satisfont aux exigences des utilisateurs. Ainsi, dans le cadre de la validation qualitative des bases de connaissances mises à disposition en tant que livrable, l'intérêt de l'usage des représentations graphiques a été démontré [Vexler et al. (2013)].

Les présents travaux s'attachent plus à la stratégie de conservation de ce niveau de qualité compte-tenu des évolutions inéluctables qu'une base de connaissance est amenée à connaître : évolution des connaissances elles-mêmes suite au Retour d'expérience (REx) ou aux efforts de R&D, mais aussi évolution de l'organisation et de ses acteurs. Les principes de similarité syntaxique ou sémantique mis en œuvre s'adaptent bien à la structuration native d'AKM.

L'exercice de ce format de restitution de travaux impose d'exprimer en langage fonctionnel opérationnel la question initiale et à présenter le résultat concret obtenu.

Nous concluons par rapport à ce retour d'expérience sur ce premier jalon que nous avons analysé comme très encourageant pour les industriels désireux de valoriser au mieux des fonds de connaissances hétérogènes dans un environnement homogène et fédérateur.

2 Cadres fonctionnel, technique et opérationnel initiaux

2.1 Origine fonctionnelle de la démarche

Le propre des grandes structures industrielles est la culture de l'écrit. Avec les contraintes imposées ou exprimées par les normes (ISO9001, ISO30401, etc.), par la législation ou la réglementation (la sûreté, le code du travail, etc.), par l'économie (les projets, les contrats, le marché, l'utilisateur final, etc.) et bien entendu par les règles de la physique (mécanique, hydraulique, chimie, métallurgie, optique, etc.), les organismes se sont dotés d'outils pour conserver leurs bonnes règles métiers, leurs meilleurs pratiques, leurs retours d'expérience, etc.

Naturellement des silos de connaissances se sont élevés dans différents services d'un même département, d'une même direction, et chacun ayant sa finalité singulière avec ses réponses dédiées. Avec le temps, les espaces de chacun de ces silos ont des intersections ou des zones de recouvrement qui s'étendent.

Puis, un jour, suite à un événement opérationnel, un manager pose la question : « pourquoi en est-on arrivé à ce stade ? », « Comment un nouvel arrivant s'y retrouve ? », « Comment est-on sûr, avec cette dispersion, de ne pas perdre la maîtrise de ce que l'on a appris ? ».

Pour résumer l'articulation du système d'information pour soutenir l'activité métier est réalisée autour de serveurs de fichiers, de système de Gestion électronique de document, de système de gestion collaborative, d'intranet métier, de système de gestion de données technique. L'organisation technique intègre les grandes phases du cycle de vie du produit (conception, production, exploitation, formation) pour simplifier de manière presque caricaturale à la complexité intrinsèque de l'activité s'ajoute celle de la sensibilité de l'information.

Fort de ce constat une réflexion sur la gestion du patrimoine de connaissance est lancée avec une cible claire de simplification à des fins d'efficacité opérationnelle tout en garantissant les usages multiples actuels et sans perdre le patrimoine aujourd'hui épars.

2.2 Structuration des bases et outil utilisé

Le méta-modèle de gestion des connaissances implanté dans AKM met en œuvre deux principes simples. Les contenus sont stockés dans des fiches ou éléments de connaissances typés par des modèles définis en fonction du besoin. L'exemple présenté dans cet article se réfère à une base de connaissance utilisant quatre modèles (Fondamental, Procédé, Fiche Technique et Fiche REx).

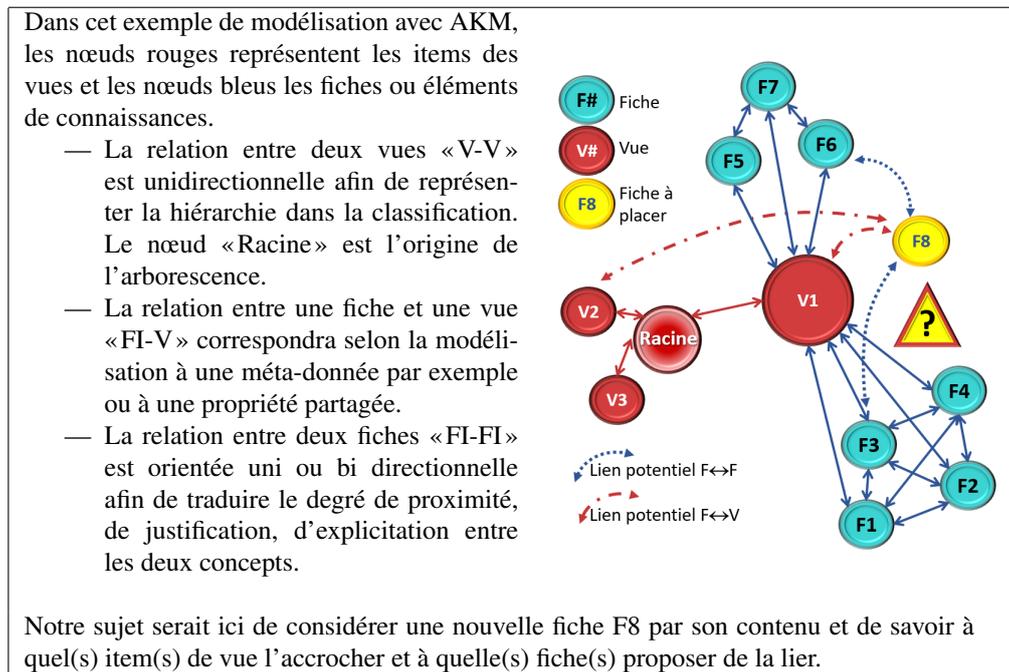


FIG. 1 – Exemple de structuration AKM - les réseaux «FI-FI» et «FI-V».

Chacune des fiches produites peut être liée à une ou plusieurs fiches via un lien bidirectionnel enregistré en base permettant la navigation de proche en proche. Ce premier réseau de liens est appelé le réseau «FI-FI». Par ailleurs, des structures arborescentes sont déclarées et jouent le rôle de support de classification pour les éléments de connaissance que l'on accroche en posant un lien. Ces structures arborescentes, que l'on appelle «Vues» dans le vocabulaire AKM, représentent les concepts métier. Elles sont susceptibles de jouer le rôle de taxonomies ou de structure de décomposition¹. Un élément de connaissance pouvant être lié à plusieurs items d'arbre via l'accrochage «multi-vues», jouera *de facto* un rôle de transversalité. Ce second réseau de liens est appelé le réseau «FI-V». Le graphe FIG. 1 page 3 illustre cette organisation. On retrouve ici les principes utilisés dans la gestion des documentations techniques modulaires

1. Lors de la conception d'une base de connaissance, on distingue bien les deux cas. Dans le cas d'une taxonomie on a une relation hiérarchique de type «Is-a» ou «SubClassOf» (ontologies) alors que dans les structures de décomposition la relation est de type «PartOf». La distinction doit être notée car les raisonnements que l'on peut mener sur ces structures sont différents.

(Data-Module et Applicabilité) mais avec un progiciel qui garantit la capacité d'évolution des structures, aussi bien des vues que des modèles. Bien entendu, les structures décrites précédemment sont autant de points d'entrée et d'accès aux contenus que de liens de navigation.

Une telle structure issue d'un processus d'explicitation au sens «recueil d'expertise» peut être vue comme un véritable «réseau sémantique», réseau sur lequel nous nous appuyons pour parfaire notre objectif d'analyse. En ce sens, ce dernier point est comparable aux approches de similarité basées sur la connaissance ou approches topologiques [Negre (2013); Elbadiry.A et Ouali.M (2015)].

2.3 Retranscription synthétique de la problématique

La question posée à l'équipe du Lab d'Ardans est comment consolider une base de connaissance distribuée car fondée sur des contenus dispersés, la fédérer autour d'un noyau élaboré sur AKM à partir d'une modélisation structurée sur une expertise humaine, la faire grandir en maintenant la consistance des éléments de connaissance entre eux (évitant les doublons) et en appuyant l'utilisateur en consultation ou en contribution dans le respect de ses prérogatives.

L'analyse de l'état de l'art sur les approches d'analyse et de comparaison de contenu a nourri la réflexion afin de renforcer le moteur AKM par un module efficient pour remplir cette fonction.

3 Processus général

Le cadre de ce travail est l'indexation automatisée de fiches de connaissances de spécialité, complétées ou non par des pièces jointes documentaires, pour laquelle on dispose de ressources sémantiques sous la forme d'une base AKM validée. Il s'agit donc dans un premier temps d'un dispositif supervisé.

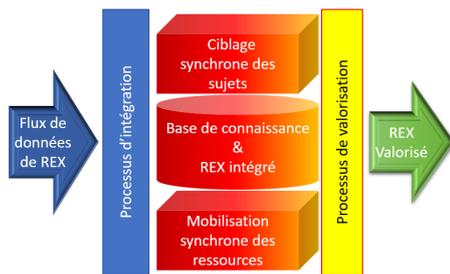
L'originalité de la démarche consiste à exploiter des recueils d'expertise et il s'agit donc là d'une manière supplémentaire de valoriser ce travail, par ailleurs utilisé de manière classique au sein des entreprises (Capitalisation & Transmission des connaissances). On trouvera dans [Luhn (1958)], dans [Elbadiry.A et Ouali.M (2015)] et dans [Desfriches Doria (2013)] les éléments de réflexion qui avalisent notre démarche.

Les fonctionnalités développées ont été mises en œuvre industriellement dans le cadre d'un «Proof Of Concept» (ci-après POC) dont l'objectif était de proposer un moyen d'intégrer des faits techniques d'exploitation (à l'origine du Retour d'Expérience ou REX) dans un référentiel commun aux différentes entités de l'entreprise et notamment les métiers². L'intérêt réside ici en deux volets organisationnels :

1. Un volet qui s'intéresse à la mobilisation du ou des métiers impliqués dans le fait technique : c'est un principe de mobilisation synchrone des ressources.
2. Un volet qui s'intéresse au ciblage du sujet à traiter et au repérage des faits techniques similaires, ce dernier point entrant dans le cadre du raisonnement par analogie.

2. Notre expérimentation revêt un intérêt d'autant plus important que l'entreprise est en charge de réaliser des produits complexes fédérant plusieurs métiers (mécanique, électronique, contrôle commande, physique et mesures, etc.).

On pourra se référer à divers articles et notamment ceux de la Foncsi (Fondation pour une culture de la sécurité industrielle) [Mortureux (2004)] pour comprendre que la gestion du REX est d'une manière générale un processus complexe à organiser. Il l'est d'autant plus que l'objet sur lequel il porte est lui-même complexe et que l'environnement d'exploitation est compliqué, avec notamment différents niveaux d'acteurs.



Chaque entité dispose alors de bases de données aux modèles plus ou moins compatibles et qui en bout de chaîne sont plus ou moins bien exploitées.

Il est donc intéressant dans ces conditions, d'exploiter au mieux la partie textuelle du fait technique avec des taxonomies prenant en compte la transversalité des sujets.

Le schéma FIG. 2 montre le dispositif avec la fonction d'intégration qui met en œuvre ces techniques de classification automatique.

FIG. 2 – Schéma du dispositif du référentiel de REX

Dans son principe, les techniques utilisées dans le POC sont relativement élémentaires et doivent être vues comme une première version de notre applicatif. L'une est directement issue de la similarité syntaxique et l'autre introduit *a minima* une similarité sémantique. Dans ce dernier cas, on considère que nous nous rapprochons plutôt d'une technique à base de corpus.

3.1 Recherche de similarité de fiches

Cette première fonctionnalité permet d'établir d'éventuels liens de fiche à fiche (réseau « FI-FI ») et réalise une véritable intégration d'une nouvelle entrée dans la base. Compte-tenu de l'organisation par modèles de cette dernière, on distingue la possibilité d'une part de relier un nouveau fait technique à un processus ou à une fiche technique (typiquement le REX de procédure et le REX de matériel) et d'autre part à un fait technique (ou REX) similaire. Le schéma FIG. 3 page 5 rend compte de cette structuration.

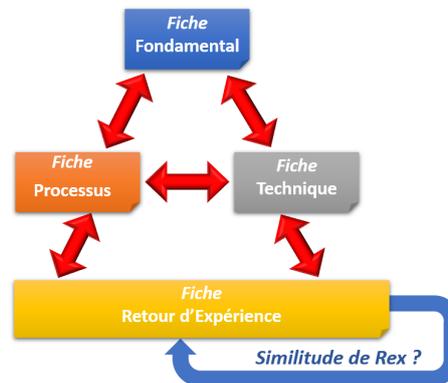


FIG. 3 – Structuration des modèles et intégration d'une nouvelle entrée REX

Le calcul de similarité est directement issu du Vector Space Model (VSM) et s'effectue selon une des trois métriques classiques : cosinus, indice de Jaccard ou distance euclidienne. Dans la partie actuellement implantée, la comparaison des fiches se fait deux à deux ce qui se justifie par le volume en général modéré des bases de connaissance³.

3. Les bases de connaissance d'expertise rassemblent quelques centaines de fiches tout au plus.

La détermination du poids se fait en utilisant une technique de normalisation entre fiches afin d'obtenir des vecteurs représentatifs indépendamment de la taille du contenu. Cette technique est celle du *tf-adapté* (Term Frequency-adjusted). Un paramétrage de poids est réalisé par l'intermédiaire de coefficients d'importance sur les différentes parties du contenu (titre, rubrique, pièce jointe : les éléments de connaissances manipulés dans nos bases).

3.2 Recherche d'items de classification

La fonction de classification sémantique est une opération originale qui consiste à déterminer un « ADN sémantique » à chaque item de la taxonomie et qui facilite la réalisation du rapprochement entre une nouvelle entrée et un ou plusieurs items.

On convient que ce principe s'approche d'une factorisation des contenus des fiches déjà rattachées à l'item qui devient alors un concept. L'approche de type « sac de mots » alias « bag of words » avec compteur brut consiste à simplement mettre en commun toutes les racines de mot de l'ensemble des fiches attachées à l'item de vue, et de traiter ce résultat comme une fiche. On aura au préalable effectué une opération de mise en exergue des termes les plus significatifs pour un item par un complément de calcul *idf* (Inverse Document Frequency).

Le calcul de similarité se produit entre la nouvelle entrée et l'« ADN sémantique » de la même manière que la similarité entre deux fiches avec normalisation de type *tf-adapté*.

3.3 Premiers résultats et suite des investigations

Les premiers essais industriels du POC⁴ ont été réalisés dans des délais tendus et la stratégie a été de valider le concept à partir de fiches fictives extraites de REX existants et volontairement dénaturés, c'est à dire moins précises ou utilisant des expressions synonymes. Nous avons pris soin notamment de supprimer certains termes fortement utilisés dans les métiers concernés. Le constat fut que ces fiches REX ont quand même été rattachées à ces métiers et aux différentes arborescences qui nous intéressaient (Système Principal, Système de Soutien, Paramètres de fonctionnement, Exigences, etc.).

De manière générale, nous avons observé pour la fonction de classification que plus un item de vue était fourni en fiches, meilleure en était la réponse. En effet, certains items n'ayant que deux ou trois fiches rattachées⁵ semblaient trop « binaires », *ite est* manquant de diversité.

Les résultats ont été suffisamment encourageant et probant pour envisager un développement plus avant vers un applicatif de type « module compagnon ».

Par ailleurs, les travaux vont changer d'échelle. Nous allons disposer d'un patrimoine de connaissance d'origine experte plus important. Les arborescences de classification seront plus riches en item et avec une sémantique induite par les réseaux de liens vers les fiches validées par les experts plus forte. Nous serons aussi en capacité de consolider et renforcer le mécanisme de rapprochement entre ce noyau d'expertise et le patrimoine «*legacy*» disséminé avec la comparaison entre différentes stratégies pour enrichir le classifieur sémantique.

4. Cette application expérimentale dans le cadre du REX est un cas particulier. Plus généralement, cette fonctionnalité est destinée à maintenir le niveau de qualité d'une base.

5. Le POC s'est effectué sur une base de connaissance restreinte ce qui explique l'existence de ces cas de figure.

4 Conclusion et perspectives

Nous avons présenté un développement adapté à une expérimentation de classification et d'intégration automatique de faits techniques (REx) au sein d'une base de connaissance. Les techniques utilisées sont jusqu'alors relativement élémentaires au regard des techniques existantes. Cependant, elles répondent déjà correctement à la problématique qui nous a été posée considérant les ressources mobilisées.

L'état de l'art que nous avons réalisé nous offre plusieurs voies d'amélioration. Ces pistes de progrès sont sereinement accessibles étant donné le matériel dont nous disposons en matière de base de connaissances et notamment la structuration avec laquelle nous construisons ces bases nous permet d'envisager toute solution qui nous semblerait pertinente parmi celle exposées dans cet article. De plus le code qui a été réalisé dans cet objectif a été conçu de manière très modulaire afin d'obtenir une adaptabilité maximale pour une efficacité optimale.

Références

- Berger, A., J. Cotton, et P. Mariot (2009). Accompagner au début du 21^{ème} siècle les organisations dans la mise en place d'une gestion des connaissances : retour d'expérience. Volume RNTI Extraction Gestion des Connaissances 2009, E-15, pp. 475–479.
- Besson, V. et A. Berger (2015). To initiate a corporate memory with a knowledge compendium : ten years of learning from experience with the ardans method. Volume RNTI Extraction Gestion des Connaissances 2015, E-28, pp. 401–412.
- Desfriches Doria, O. (2013). *Faceted Classification for Knowledge Management in specific trades, Method for the development of FolkFaceted Classifications*. Theses, Conservatoire national des arts et métiers - CNAM.
- Elbadiry, A. B. et Ouali, M. (2015). Etude comparative des méthodes d'analyse de similarité des défaillances de système aéronautiques. Québec.
- Luhn, H. P. (1958). A business intelligence system. Volume 2, pp. 314–319.
- Mortureux, Y. (2004). Comparaison de textes : quelques approches... *Techniques de l'Ingénieur, Réf: SE1040 v1*.
- Negre, E. (2013). Comparaison de textes : quelques approches... *Hal-00874280*.
- Vexler, F., A. Berger, J. Cotton, et A. Belloni (2013). Eléments d'appréciation et d'analyse d'une base de connaissance : l'expérience industrielle d'Ardans. Volume Actes Atelier AIDE EGC'2013, pp. 59–72.

Summary

The continuous supply of knowledge bases by the contributors is an essential point of the interest and the durability of these devices within the organizations. Beyond the production effort, comes the need for integration of content within an existing guaranteeing the consistency of the device. The latter can be complex, especially since it is often multidisciplinary.

The knowledge bases are built using the Ardans Knowledge Maker®(AKM) tool, and therefore strongly structured in terms of templates and classification trees (or views), the usual

method relies on the expertise of the contributor. This mode of operation can however be faulted, especially during occasional contributions or when it is desired to integrate content in an automated process.

Coming back to an automatic classification problem, Ardans uses vector semantics to implement a knowledge-driven classification process. This approach is an important aid in maintaining consistency.

After a reminder of the main techniques of the field, this article discusses the realization made and its use in a recent «Proof Of Concept» for the automatic classification of new feedback (REX). This work comes from industry and twenty years of knowledge engineering operations carried out by Ardans in France and Europe. They foreshadow the next use of AKM's companion module called «Semantic Analysis».