# De la page blanche à la boîte noire : quand le TALN devient éminence grise

Richard Launay, Céline Raynal, Jean-Marie Rousseau

Institut pour la Maîtrise des Risques
Sûreté de Fonctionnement - Management - Cindyniques

# De la page blanche à la boîte noire : quand le TALN devient éminence grise

# Looking inside the black box : is NLP the next mastermind ?

Richard LAUNAY
Institut de Radioprotection et de Sûreté Nucléaire (IRSN)
Fontenay-aux-Roses, France
richard.launay@irsn.fr

Céline RAYNAL
Safety Data - OmniContact
Paris, France
celine.raynal@safety-data.com

Jean-Marie ROUSSEAU
Institut de Radioprotection et de Sûreté Nucléaire (IRSN)
Fontenay-aux-Roses, France
jean-marie.rousseau@irsn.fr

*Résumé*

Les exploitants nucléaires produisent des comptes rendus d'analyse de tout événement significatif. Un démonstrateur mettant en œuvre des algorithmes de traitement automatique du langage naturel a été testé. Les résultats sont prometteurs, mais questionnent sur la place qu'ils peuvent prendre au regard des transitions qui s'opèrent dans les pratiques d'expertise.

*Abstract*

Nuclear plant operators must produce analysis reports for any significant event. A demonstrator implementing natural language processing algorithms was tested on them. The results are promising but question the place they can have considering the transitions that occur in the practices of expertise.

.

***Mots clefs— Retour d'expérience, Traitement Automatique des Langues (TAL), Apprentissage supervisé, Organisation, Transition***

***Keywords— User feedback, Natural Language Processing (NLP), Supervised machine learning, Organization, Transition***

## I. INTRODUCTION

L'institut de radioprotection et de sûreté nucléaire (IRSN) collecte tous les comptes rendus d'événements significatifs (CRES) déclarés à l'autorité de sûreté par les exploitants d'installations nucléaires de base (INB). Ces documents constituent la « matière première » du traitement du retour d'expérience (REX) opéré par l'Institut dans le cadre de l'amélioration de la sûreté des installations nucléaires.

Le traitement de ces CRES par l'IRSN revient en partie à les catégoriser afin d'identifier des enseignements pertinents vis-à-vis de la maîtrise des risques. Pour ce faire, l'IRSN utilise une grille de lecture des événements reposant sur un modèle d'analyse transverse (A2N-T). Le travail de catégorisation repose ainsi sur une analyse linguistique du texte disponible. Il est réalisé par des experts qui doivent comprendre le contexte opérationnel (technique et organisationnel) dans lequel l'événement survient afin de « décrypter l'histoire » racontée. Ils doivent également maîtriser le modèle d'analyse établi en amont afin « d'encoder les enseignements pertinents » à retenir de cette histoire.

L'IRSN a décidé de s'appuyer sur les techniques de traitement automatique du langage naturel (TALN) pour aider les experts dans cette tâche et ainsi recentrer leur travail (moins de temps de lecture, plus de temps d'analyse) sur les enseignements à tirer de l'ensemble des événements déclarés.

Le présent article vise à présenter les opportunités, les risques et les modifications notables dans l'expertise de l'IRSN, induites par l'utilisation des techniques de TALN. Dans un premier temps nous rappellerons les bases de la méthode A2N-T puis son intégration dans un modèle algorithmique de TALN. Sur la base d'une étude de cas, nous présenterons les apports et limites de ces outils d'intelligence artificielle, notamment pour le travail d'expertise. Nous discuterons ainsi de la place donnée à ces outils dans l'activité qu'ils supportent.

## II. LE TRAITEMENT DU REX A L'IRSN

Les exploitants des installations nucléaires de base (INB) ont l'obligation réglementaire de déclarer aux autorités les événements non souhaités auxquels ils sont confrontés et qui peuvent avoir des impacts sur la sûreté, la radioprotection des travailleurs, la protection de l'environnement. Chaque événement fait l'objet d'un rapport d'analyse d'une vingtaine de pages en moyenne (CRES) réalisé par l'exploitant concerné. L'IRSN stocke ces déclarations et ces analyses dans une base de données, afin de réaliser une analyse « de second niveau » des CRES transmis par les exploitants, soit environ 1 300 par an.

### *A. Présentation des données*

Le récit de l'événement non souhaité, élaboré par les exploitants, constitue la « matière première » du retour d'expérience. Mais au-delà de ce récit, ce sont avant tout des milliers de « mots » qui constituent les « données textuelles » au sens de l'utilisation que nous allons présenter par la suite.

On dispose ainsi de données qui peuvent être décrites sous la forme de milliers de mots, d'acronymes, de valeurs alphanumériques, de schémas, de tableaux, de graphes, de formules chimiques ou mathématiques, de photos, etc. Tous ces éléments constituent les « données » et participent à la compréhension de l'événement et des enseignements qu'il porte.

Les CRES ont une structure relativement commune, la trame générale ayant été décrite par l'autorité de sûreté et permettant ainsi de spécifier les éléments de récit qui doivent être présentés : la chronologie ; les causes de l'événement ; les conséquences réelles et potentielles ; les mesures préventives et correctives. Cette structuration narrative quasi « identique » des textes revêt une importance considérable pour le traitement par des algorithmes. En effet, plus la structure est homogène sur l'ensemble des CRES, plus les traitements proposés (de comparaison, notamment) pourront être pertinents.

On peut ainsi identifier dans ces documents plusieurs types d'informations primordiales. On aura dans un premier groupe toutes les données qui sont propres à l'identification générale de l'événement, que l'on va appeler les « métadonnées ». Ce sont des données pré-structurées de type « attribut/valeur », les valeurs étant comprises dans des listes finies. Il pourra s'agir par exemple, de l'exploitant (CEA, EDF, ORANO, ANDRA, etc.), du numéro d'INB, du site, de l'état de l'installation (exploitation, à l'arrêt, etc.), des matériels impliqués, du critère de déclaration concerné, etc. Dans un second groupe qui correspond au cœur du récit, on va trouver des données textuelles, non structurées, qui sont le résultat d'une analyse a posteriori d'un événement non souhaité. Cette analyse de l'événement se traduit par un récit (destiné à des lecteurs internes ou externes) qui permet normalement d'identifier les séquences de faits (habituels et inhabituels) qui interviennent dans le déroulement d'un scénario événementiel.

Ces données ne constituent pas à proprement parlé des « Big Data », mais présentent néanmoins un certain « volume ». A terme, le projet est d'ajouter à ces CRES toutes les « fiches d'écarts » produites par les exploitants, tous les comptes rendus d'inspections internes et externes et tous les documents relatifs aux incidents, accidents hors de la sphère nucléaire, on pourra alors avoir affaire à des données complémentaires qu'il s'agira de mettre en perspective les unes par rapport aux autres.

Les récits à la disposition de l'IRSN le sont sous la forme de documents de type « pdf » nécessitant une « océrisation ».

### *B. Les utilisations courantes des données dans le REX*

Aujourd'hui, les comptes rendus d'événements sont compilés dans des bases de données informatiques dans lesquelles on retrouve la décomposition décrite ci-dessus (métadonnées et récits). On y trouve également des indexations de mots clefs prédéterminés qui correspondent à un modèle de « catégorisation » mis en place historiquement, suivant une logique « indicateurs » et permettant des recherches « critérisées ».

L'IRSN utilise ces documents au travers des outils aujourd'hui à disposition, mais surtout au travers du prisme constitué par l'expertise de ses chargés d'affaires. L'analyse de ces comptes rendus est portée par chaque chargé d'affaire de l'installation désignée. Il peut ainsi, au regard de sa connaissance de l'installation, du contexte, de sa situation administrative, interagir avec ses homologues exploitants ou l'autorité de sûreté pour affiner l'analyse, accompagner les actions de contrôle, etc. Mais, pour chaque chargé d'affaires, ces analyses sont autant d'informations lui permettant d'identifier les failles potentielles de l'installation dont il a la charge et d'orienter les actions d'amélioration requises. L'étude des CRES permet également d'identifier plus globalement des problématiques récurrentes, des tendances sur une installation, sur un ou plusieurs sites, voire un ou plusieurs exploitants.

Historiquement, ces données étaient essentiellement utilisées pour produire des données quantitatives mettant en exergue principalement les informations issues des « métadonnées ». Le nombre d'événements déclarés est ainsi comptabilisé par exploitant, par installation, sur l'année écoulée en comparaison avec les années précédentes en considérant que tout événement déclaré a le même poids et est porteur d'un enseignement spécifique, implicite. Ce nombre d'événements peut également être croisé avec des thématiques prédéterminées, comme par exemple le nombre d'événements liés à la problématique de la radioprotection sur une année en comparaison aux années précédentes. Il est aussi possible d'essayer de tirer des enseignements structurels ou organisationnels à partir d'un nombre croissant de contrôles et essais périodiques (CEP) non réalisés ou d'opérations de maintenance conduisant à des non-conformités. Cette forme de REX basée sur l'analyse d'indicateurs de type « tableaux de bord » ne peut suffire pour expliquer la réalité de situations événementielles complexes. L'analyse fouillée de ces situations repose notamment sur : la qualité des CRES produits par les exploitants ; la connaissance des installations des chargés d'affaires de l'IRSN et celle des modèles de maîtrise des risques qui permet d'identifier les facteurs potentiels de dysfonctionnements (source du REX industriel, connaissances des accidents majeurs…).

### *C. Le modèle A2N-T et son utilisation*

L'IRSN s'est engagé fin 2014 dans une réflexion pour améliorer le traitement des événements significatifs déclarés afin de renforcer la transversalité et la pertinence des analyses [1]. Cette réflexion a abouti à formaliser des éléments méthodologiques permettant de soutenir quatre fonctions principales pour le traitement du REX :

- porter un jugement critique sur le CRES, afin de fiabiliser les données d'entrée de nos propres analyses ;
- suivre l'évolution d'une variable particulière à travers une démarche de type « relevé d'indicateurs », notamment pour surveiller l'évolution d'un phénomène pré-identifié ;
- repérer des récurrences sur un ensemble d'événements et porter des alertes sur un type d'événement, un site particulier, une activité particulière, une population particulière d'intervenants, un phénomène (causal)

particulier, permettant de questionner la pertinence des actions correctives ;

- favoriser les actions proactives de l'Institut à partir de tendances observables (précurseur d'une dégradation, potentiel générique, etc.) tirées de ces analyses.

Pour ce faire, un « modèle de maîtrise des risques » a été développé (listant les facteurs qui contribuent à la maîtrise attendue). Inscrit dans le cadre conceptuel de la défense en profondeur [2], il constitue l'ossature de la lecture transverse des récits à analyser. Le modèle est structuré par la notion « d'activité » ; l'activité étant « *ce qui est fait* » par un ensemble d'acteurs organisés. La maîtrise de ces activités passe par la mise en place de dispositions qui relèvent de :

- la planification : les activités sont ordonnées dans le temps, mobilisant des ressources temporelles (durée), techniques et humaines spécifiques ;
- la préparation : les objectifs, les dispositions de maîtrise des risques, les moyens pour les réaliser sont explicités, compris et connus par les différents acteurs ;
- la réalisation : les activités sont effectuées par des intervenants mobilisant des ressources individuelles ou collectives, cognitives (intentions, décisions) ou physiques (actions) ;
- l'encadrement technico-organisationnel : un contexte sociotechnique (environnement technique, humain, organisationnel, managérial, économique et réglementaire) influence la manière dont la planification, la préparation et la réalisation se déroulent.

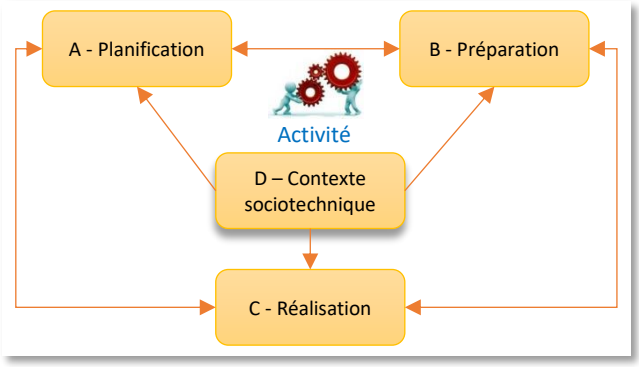


Fig. 1 : Structure du modèle A2N-T

Au sein de ces quatre blocs, des « lignes de défense » (LdD) sont censées assurer des fonctions spécifiques de maîtrise des risques. Trente lignes de défense sont ainsi explicitées : trois relatives à la planification par exemple la prise en compte de l'adéquation de la charge et des ressources dans la planification des opérations d'exploitation (A2), huit relatives à la préparation, par exemple la ligne de défense B4 relative à la prise en compte du REX en phase de préparation, cinq relatives à la réalisation, par exemple C3 qui traduit les actions efficaces de contrôle et vérification des opérations d'exploitation et quatorze qui définissent le contexte sociotechnique par exemple D5 qui traduit un référentiel documentaire d'exploitation adapté à son usage.

L'analyse des CRES consiste alors à rechercher, dans le récit produit par l'exploitant, la contribution défaillante ou efficace de ces lignes de défense dans le scénario événementiel. En constatant le succès ou l'échec des LdD, il s'agit d'identifier et de comprendre les phénomènes qui perturbent/facilitent la mise en œuvre des dispositions de maîtrise des risques proposés par le système sociotechnique. Bien sûr, la profondeur et l'objectivité de l'analyse restituée par l'exploitant dans le CRES, ainsi que la connaissance que l'analyste a du référentiel de l'exploitant, sont des sources de biais potentiels dans cet exercice de compréhension, à défaut d'échanges directs entre les différents acteurs.

Les CRES sont ainsi tous « encodés » avec ce modèle. Ce « codage manuel » revient à associer à chaque événement des caractéristiques complémentaires (« signature » constituée de la liste des LdD défaillantes ou efficaces) aux descripteurs plus factuels (date, site, fonction de sûreté impactée, etc.). Le « codage » d'un ensemble d'événements permet de tirer des enseignements divers sur, notamment :

- la pertinence du modèle de maîtrise des risques conçu et mis en œuvre par les exploitants pour opérer en toute sûreté ;
- le « taux de défaillance/succès » pour chacune des LdD au sein d'un échantillon particulier d'événements ;
- des « profils comparatifs de défaillances » entre deux échantillons constitués à partir de critères variés (n installations, n années, n typologies d'événements, etc.) ;
- la dépendance statistique entre les lignes de défense, à travers l'explicitation des liens de fragilisation entre LdD défaillantes [3] ;
- la pertinence des actions correctives retenues par l'exploitant par comparaison avec les LdD identifiées comme défaillantes : actions couvrant de manière adéquate les défaillances identifiées / actions hors sujet (ne portant sur aucune défaillance identifiée) / action manquante (défaillance non couverte par une action corrective).

### *D. Des apports et des limites*

Les fondements de ce modèle et sa déclinaison opératoire pour l'analyse des événements constituent l'outil conceptuel et opérationnel dont l'IRSN s'est doté pour proposer une base commune de questionnement pour l'ensemble de ses chargés d'affaires intervenants dans les activités d'expertise. L'approche « systémique » retenue permet également d'analyser tous les événements de tous les exploitants et de tous les types d'installations sous le même angle, ce qui constitue un élément primordial pour assurer une vision transverse des risques et des enseignements tirés du REX. Cela permet également de franchir la barrière de l'analyse des événements spécifiques à la sphère nucléaire pour enrichir le REX de tous les événements connus et disponibles sous la forme d'un récit.

Mais ce « codage » à hauteur d'environ 1 300 événements par an est extrêmement chronophage. S'il est réalisé par une petite équipe spécialisée, il peut revêtir une forme « rébarbative » qui place alors l'expert non plus dans un schéma d'analyse réflexive, mais dans une forme de travail routinier hyper-spécialisé (« à la chaîne ») qui peut nuire à la réflexion, tant sur le modèle que sur le produit rendu. Une organisation centralisée de l'activité risque d'enfermer chaque codeur dans sa propre représentation du modèle A2N-T, par

rapport à une organisation plus « distribuée ». Le risque est alors de passer d'une approche systémique à une vision anthropocentrée des événements qui nuira indubitablement à la qualité du « codage », mais aussi plus globalement biaisera les analyses et les résultats. Le recours à une certaine forme d'automatisation est dès lors envisageable.

## III. LES OUTILS D'INTELLIGENCE ARTIFICIELLE

Après ce travail essentiel sur le modèle de maîtrise des risques, l'IRSN a mené plusieurs expérimentations au travers de POC (*Proof of Concept*), et notamment un sur les technologies de traitement automatique du langage naturel (TALN). Cela s'est concrétisé avec la mise en essai de l'outil *PLUS* développé par l'entreprise Safety Data.

### A. *Description de l'outil PLUS*

Safety Data (groupe OmniContact) développe depuis plusieurs années l'application web *PLUS* (pour *Processing Language Upgrades Safety*) dont l'objectif est de faciliter, voire de permettre, l'exploration et l'exploitation de bases de données textuelles disponibles dans les organisations, et plus spécifiquement celles confrontées à la gestion des risques.

En effet, le volume croissant de données collectées et stockées numériquement dans les organisations pose la question de leur traitement. Dans un grand nombre de cas, et notamment celui du retour d'expérience, ces données sont largement composées de textes. Or, s'il est courant et aisé d'interroger les données structurées (listes de valeurs, métriques, booléens, etc.), il est plus difficile d'interroger efficacement les données « non structurées » (les textes), par définition très hétérogènes. En effet, on observe que pour un même événement, chacun des acteurs présents relate les faits selon ses propres mots, différents de ceux des autres.

L'objectif de *PLUS* est de permettre d'interroger ces données textuelles extrêmement variées. Pour ce faire, le cœur de l'application consiste en une analyse linguistique fine, puissante et adaptée à la langue de spécialité utilisée dans les documents traités. En effet, bien que les retours d'expérience soient toujours rédigés en « langue naturelle », cette langue comporte des spécificités propres au domaine : celles du nucléaire ne sont pas les mêmes que celles présentes dans l'énergie hydraulique ou dans l'aéronautique par exemple. Ainsi, « AAR » renvoie à un « Arrêt Automatique du Réacteur » pour l'IRSN tandis qu'il correspond aux « Archives Audiovisuelles de la Recherche » ou « Anévrisme de l'Artère Rénale » dans d'autres domaines.

Une fois les textes analysés linguistiquement, une version structurée de ces textes est disponible et c'est la base des textes structurés qui va alors faire l'objet des traitements ultérieurs. Ainsi, les requêtes définies par les utilisateurs dans le moteur de recherche de *PLUS* vont elles-mêmes être analysées linguistiquement et leur version structurée confrontée à la base des textes structurés pour renvoyer les documents les plus pertinents. Cette mécanique permet ainsi, par exemple, d'obtenir les mêmes résultats que l'on cherche « arrêt ventilateur » ou « ventilation arrêtée ». Il en va de même pour toutes les fonctionnalités de TAL disponibles dans *PLUS* : l'analyse de similarité textuelle, qui consiste à comparer les textes de la base de données les uns aux autres afin de trouver ceux qui partagent le plus d'informations, la catégorisation automatique et la création de dimensions. Des précisions sur la démarche scientifique et les modèles utilisés sont disponibles dans [5].

La création de dimensions utilisées par l'IRSN est présentée ci-après en détail, à partir d'exemples précis[1]. Précisons toutefois dès à présent qu'elle consiste à modéliser l'expression linguistique d'une thématique spécifique afin de s'affranchir de la recherche par un ensemble de mots clefs fixant le contour sémantique de la thématique considérée. Il est donc crucial que l'analyse linguistique faite en amont soit la plus pertinente possible. Une fois le modèle de la dimension appris sur un corpus d'apprentissage défini par l'expert, il est automatiquement appliqué à tous les documents de la base de données et un poids est attribué à chacun d'entre eux : plus la dimension est présente dans le document, plus le poids associé tend vers 1, inversement si la dimension est absente, son poids tend vers 0. Cette représentation des résultats permet une recherche sur une métrique linguistique et non plus sur le texte en tant que tel : recherche bien plus puissante fournissant des résultats plus riches et plus fins.

### B. *Le modèle A2N-T dans PLUS*

L'outil *PLUS* a été testé au travers de ses différentes fonctions que sont l'analyse de similarité, la catégorisation et la construction de dimensions. Après plusieurs phases de tests, il a semblé que la construction de dimensions était particulièrement intéressante pour modéliser les différentes lignes de défense du modèle A2N-T.

Il faut noter que l'utilisation de l'analyse par les « dimensions » s'est avérée pertinente parce que nous disposions d'un modèle conceptuel et systémique de maîtrise des risques (A2N-T). La validation des résultats était également rendue possible dans la mesure où nous disposions de plus de 6 000 CRES codés manuellement. Ainsi, nous savions quoi rechercher dans les textes et comment évaluer la pertinence des résultats produits. Nous verrons cependant dans un second temps que même en partant « de zéro », il est possible de créer des dimensions pertinentes [III-C]. De ce fait, l'utilisation d'une technique d'intelligence artificielle mettant en œuvre l'apprentissage supervisé a semblé particulièrement adaptée à notre besoin. Pour chaque ligne de défense du modèle, nous avons pu construire les dimensions associées. La méthode pour construire ces dimensions se déroule en trois étapes, comme indiqué dans la figure suivante :

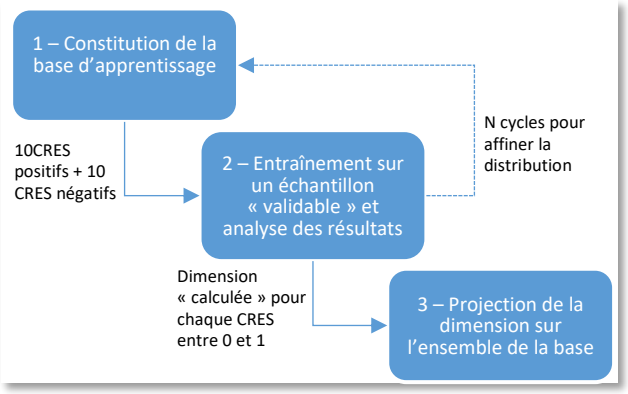


Fig.2 : Méthode de construction d'une dimension avec *PLUS*

[1] Nous renvoyons à [4] pour la présentation d'une première réalisation autour des dimensions dans le domaine de la santé.

Pour chaque ligne de défense, nous avons dans un premier temps isolé des CRES « positifs », c'est-à-dire pour lesquels nous avions identifié cette ligne de défense comme défaillante, et des CRES « négatifs », c'est-à-dire n'ayant pas été codés défaillants (ni efficaces) sur cette même ligne de défense. Ces documents ont été recherchés et extraits chez tous les exploitants pour s'affranchir autant que faire se peut d'un biais de codage lié à un exploitant qui mobiliserait un vocabulaire trop spécifique. Ces CRES ont été validés par une tierce personne pour tenter de supprimer un biais de codage/codeur et vérifier la qualité du texte transféré dans *PLUS* (problème de l'océrisation des documents PDF). L'apprentissage débute en général avec une dizaine de documents dans chaque catégorie (positifs/négatifs).

Une fois l'apprentissage réalisé (« entraînement » de l'algorithme), les dimensions sont « projetées » avec une valeur décimale comprise entre 0 et 1. La médiane de 0,5 représente les textes que l'algorithme n'a pas pu distribuer. Elle correspond ainsi à une position indéterminée vis-à-vis du corpus d'apprentissage.

Dans un premier temps, cette distribution est analysée sur un corpus de référence, c'est-à-dire les CRES codés manuellement avec la LdD modélisée par la dimension. Dans l'exemple ci-dessous, ce sont les CRES codés manuellement avec la ligne de défense défaillante « C1 – Décision inappropriée en temps réel » qui sont considérés, soit 1 012 CRES dans lesquels des traces de défaillance de cette ligne de défense ont été relevées.

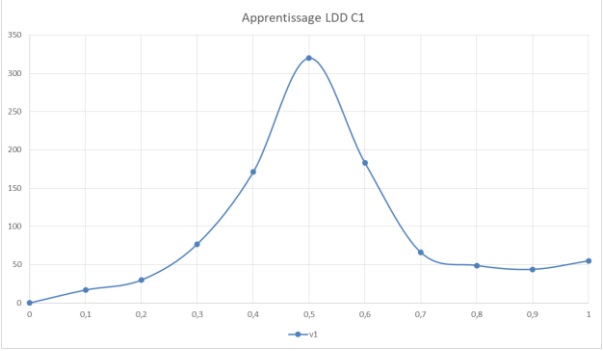


Fig.3 : Première courbe d'apprentissage C1

La figure n°3 présente la répartition des valeurs dans l'intervalle [0 – 1]. Le suivi des résultats est apparu assez rapidement essentiel pour évaluer visuellement l'évolution de l'apprentissage. Logiquement, nous devrions avoir les 1 012 CRES avec une valeur proche de 1. Or, la gaussienne obtenue après le premier apprentissage indique qu'une majorité des CRES sont calculés comme indéterminés. Il est donc nécessaire de poursuivre l'entraînement. Plusieurs boucles itératives sont nécessaires et réalisées pour affiner la distribution. Dans une première phase, il s'agit à partir du corpus des CRES codés manuellement (positivement) de forcer l'apprentissage en intégrant pas à pas les CRES par pas de 0,2. Ce qui revient *in fine* à réaliser au moins 8 entraînements ce qui nous amène à un corpus d'apprentissage contenant une trentaine de CRES positifs et négatifs. On obtient alors ce type de graphes :

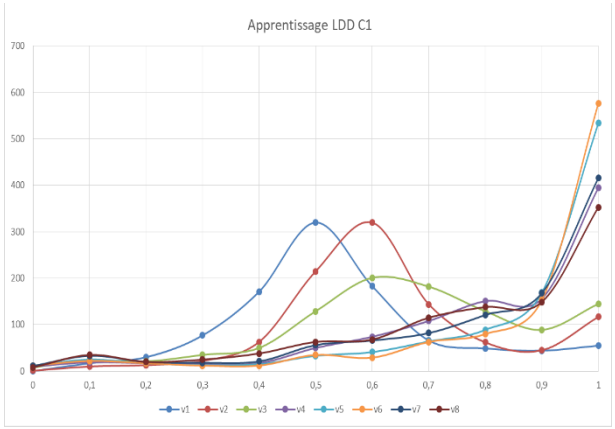


Fig.4 : Courbes d'apprentissage LdD C1

La figure n°4 présente les graphes d'évolutions de l'apprentissage avec dans un premier temps une gaussienne comprise entre 0,4 et 0,7 (entraînements 1, 2 et 3) puis, au fur et à mesure des entraînements, des CRES qui se distribuent vers les valeurs comprises entre 0,8 et 1, tendant ainsi vers les résultats du « codage manuel ». Ces résultats sont présentés dans le tableau n°1 qui présente l'évolution de la répartition des CRES dans chaque intervalle de confiance pour chaque version d'apprentissage.

| | v1 | v2 | v3 | v4 | v5 | v6 | v7 | v8 |
|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 1 | 8 | 10 | 12 | 12 | 11 | 8 |
| **0,1** | 17 | 10 | 23 | 19 | 25 | 22 | 33 | 35 |
| **0,2** | 30 | 13 | 21 | 17 | 17 | 17 | 20 | 20 |
| **0,3** | 77 | 22 | 35 | 15 | 13 | 12 | 18 | 25 |
| **0,4** | 171 | 63 | 50 | 17 | 15 | 12 | 21 | 38 |
| **0,5** | 320 | 215 | 129 | 50 | 33 | 35 | 56 | 63 |
| **0,6** | 183 | 320 | 201 | 74 | 41 | 29 | 66 | 68 |
| **0,7** | 66 | 144 | 182 | 108 | 64 | 63 | 82 | 115 |
| **0,8** | 49 | 62 | 129 | 151 | 89 | 80 | 121 | 138 |
| **0,9** | 44 | 45 | 89 | 156 | 169 | 153 | 168 | 149 |
| **1** | 55 | 117 | 145 | 395 | 534 | 577 | 416 | 353 |

Tab.1 : Répartition des CRES (version d'apprentissage 1 à 8) dans les intervalles de confiance

D'une façon générale, tous les graphes d'apprentissage suivent cette tendance ; celle-ci est d'autant plus marquée que le nombre de CRES associés à la ligne de défense est important (> 800). Il demeure quelques difficultés pour des dimensions qui modélisent des lignes de défense peu codées (base d'apprentissage inférieure à 100 documents).

Une fois l'apprentissage stabilisé, il a paru intéressant d'analyser chacune des dimensions, non plus sur son corpus positif (ensemble d'événements dans lequel on s'attend à trouver la dimension avec une valeur proche de 1), mais sur l'ensemble de la base. L'idée est alors d'identifier la variation entre le codage manuel « expert » et le codage « algorithmique ». Cela peut également révéler des défauts de codage manuel pouvant ainsi potentiellement orienter l'utilisation de l'outil non plus comme une aide à la décision mais vers une forme de super contrôleur du codage manuel.

Cette analyse a permis d'identifier des erreurs de codage qui sont essentiellement des oublis ou des évolutions de pratiques de codage apparues au cours du projet (période de stabilisation et d'appropriation du modèle A2N-T). Mais cela a également permis d'identifier une surestimation du codage algorithmique. La dimension affectant une valeur comprise

entre 0,8 et 1 à de nombreux CRES (environ 20%) constituant ainsi des faux positifs.

Une seconde phase d'apprentissage a alors été envisagée en répartissant des CRES de l'ensemble de la base dont la valeur est comprise entre 0,4 et 0,6 dans les positifs ou négatifs et ce pour chacune des dimensions. On tente ainsi d'affiner l'apprentissage mais cette fois en utilisant les réponses algorithmiques sur la totalité de la base. Les faux positifs tendent alors à refluer pour rééquilibrer l'ensemble du codage. Il faut noter que la performance du codage algorithmique est variable d'une dimension à l'autre en fonction de la nature de la ligne de défense modélisée. Certaines LdD sont plus « conceptuelles » que d'autres en ce sens qu'elles se réfèrent à des marqueurs linguistiques peu discriminants : un nombre élevé de mots à forte occurrence (vocables banals du corpus textuel) rendant la distribution moins tranchée (plus gaussienne). En comparaison, les dimensions portant sur des lignes de défense à composantes plus « descriptives », donc linguistiquement plus marquées (même variées), produisent des résultats plus conformes à l'attendu (taux élevé de reconnaissance de la dimension, sans erreur).

### *C. Un exemple d'utilisation sur un thème d'expertise*

L'apprentissage supervisé semble prometteur pour s'orienter vers un codage automatique mais peut-on l'utiliser d'une autre façon ? Par exemple, on peut s'intéresser, non pas directement à une ligne de défense, mais à une problématique spécifique, par exemple, la contamination corporelle de salariés œuvrant dans l'industrie nucléaire. Sujet d'importance en termes de conséquences potentielles mais également au regard de sa possible extension aux autres domaines industriels qui vont de la contamination chimique à la contamination biologique.

En lançant la requête (textuelle) « contamination corporelle », 61 comptes rendus sont extraits sur les 6200 documents contenus dans la base. La recherche ramène par exemple ce type d'information :

- « *… une contamination corporelle sur le genou droit est révélée* »
- « *Contamination corporelle superficielle d'un opérateur au niveau de la paume de la main au cours d'une manipulation en boîte à gants.* »
- « *Une contamination corporelle externe est détectée à hauteur du nez…* ».

Mais cette recherche « simple » ne permet pas de discriminer « la présence de contamination corporelle » de « l'absence de « contamination corporelle », signifiée par les phrases du type :

- « *l'investigation a montré l'absence de contamination corporelle* ».
- « *Les contrôles radiologiques ont révélé une contamination de l'extérieur de la cartouche et du masque de l'opérateur ainsi que l'absence de contamination corporelle* ».

Le moteur de recherche de *PLUS* permet de pallier ce problème en utilisant des filtres pour retenir les textes qui contiennent « contamination corporelle » et exclure ceux qui contiennent « absence de contamination corporelle ». Le résultat passe ainsi de 61 à 51 CRES. Mais si ce mode de recherche réduit le « bruit », il occulte les autres formes syntaxiques possibles, comme par exemple :

- « *pas de contamination corporelle* »,
- « *pas révélé de contamination corporelle* »,
- « *pas relevé de contamination corporelle* ».

De la même façon, une « *contamination de la main droite* » est équivalente (linguistiquement) à une « *contamination corporelle* » (ce qui n'est pas le cas, médicalement parlant). On peut également trouver ce type d'information : « *La sonde passée au niveau du visage révèle la présence de contamination concentrée au niveau de la bouche et de la barbe (environ 60 c/s en sonde bêta)* » sans que les mots « *contamination corporelle* » ne soient utilisés, le résultat échappant ainsi à la fouille de texte.

On voit par ces quelques exemples que les moteurs de recherche textuels (fouille de textes) sont efficaces à condition de réaliser *a priori* des constructions qui permettent des recherches multiples et très variées, ce qui s'avère pratiquement impossible tant cela devient complexe. En effet, en considérant l'exemple ci-dessus, il paraît évident que « *bouche* » est une partie du corps et que l'on souhaiterait que la recherche du terme « c*orporel* » permette de trouver toutes les parties du corps. Cela étant, on se rend très vite compte – notamment dans un contexte industriel – que « *bouche* » peut également faire référence à une entrée (comme dans « *bouche d'aération* », par exemple). Il en va de même pour « *pied* » (« *contrôleur mains pieds* »), « *main* » (« *courante* »), etc.

Dans les outils d'archivage et de recherche courants, le lecteur (analyste) est la parade. En effet, ces textes sont, après lecture par un expert, indexés et associés à un ou plusieurs mots clefs ou à une idée. Par exemple, le mot clef « *contamination* » regroupe des CRES qui ne sont pas spécifiques à la contamination d'un opérateur mais qui traitent de la contamination interne et externe d'opérateurs mais aussi de matériels et objets du procédé, de l'environnement ou de risques potentiels jugés dimensionnant ou intéressants. Cela implique que tous les CRES doivent être lus avec attention et indexés à des mots volontairement peu nombreux ou vagues pour faciliter l'indexation. Cette façon de procéder oblige à figer les mots clefs ou la structure *a priori* des données, car toute modification implique la ré-indexation de la totalité de la base. *In fine*, on préfère construire de grandes catégories (incendie, conditions météorologiques, contrôle commande, maintenance, etc.) dans lesquelles, il est possible de se replonger si besoin, pour une analyse plus fine. C'est une façon de « diviser (les données) pour régner (sur le temps d'analyse) ».

Pour contourner cet écueil, nous avons utilisé le procédé décrit précédemment, en construisant une dimension spécifique à la « *contamination corporelle* ». Il n'existait pas de corpus positif et négatif. Le premier travail a consisté à les construire à partir de recherches successives de mots clefs, supposés représentatifs de cette thématique (contour sémantique). L'entraînement a débuté avec sept documents positifs et négatifs. La supervision de l'apprentissage a été réalisée en utilisant le corpus de référence construit sur la base de documents contenant quelques mots clefs associés à la « *contamination* » mais pas obligatoirement corporelle. La base de référence est ainsi constituée de 1 583 CRES. Les résultats du premier apprentissage sont donnés dans la figure n°5 :

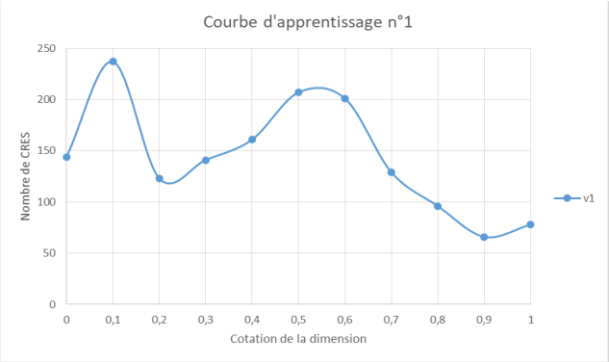


Fig.5 : Courbe d'apprentissage n°1 « contamination corporelle »

L'apprentissage s'est poursuivi comme indiqué dans le paragraphe [III.B] pour obtenir *in fine* les courbes d'apprentissage présentées dans la figure n°6.

Celle-ci montre que l'apprentissage fonctionne car à chaque entraînement, les CRES se distribuent dans les valeurs comprises dans l'intervalle [0 – 0,2] ce qui est attendu, puisque l'échantillon de référence contient relativement peu de CRES faisant référence à une contamination corporelle d'opérateurs. A l'inverse, l'apprentissage n°4 nous indique qu'il ne reste plus que 46 CRES dont le poids de la dimension est compris dans l'intervalle [0,8 – 1] c'est-à-dire qui ont un rapport avec « *la contamination corporelle* » d'un opérateur. Cette dimension relative aux « *contaminations corporelles* », dans sa version définitive, a été réalisée en utilisant 14 CRES positifs et 14 CRES négatifs et évaluée sur une collection de 1 583 CRES.

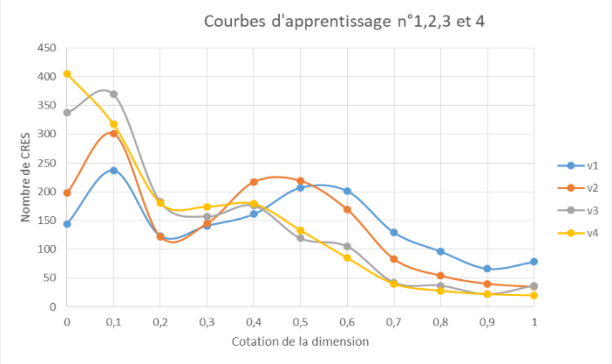


Fig.6 : Courbes d'apprentissage 1, 2, 3 et 4

### *D. Résultats et discussions*

Cette dimension s'est vue projetée à l'ensemble de la base (6 200 CRES). On obtient alors 46 CRES compris dans l'intervalle de confiance [0,8 – 1], ce qui permet à l'expert de se focaliser uniquement sur ces derniers, auxquels on peut soustraire les 14 CRES ayant servi à la construction de la dimension. L'expert peut donc se focaliser sur 32 CRES.

Il apparaît que sur ces trente-deux CRES, cinq sont des faux « positifs » (sûrement améliorable en affinant l'apprentissage, c'est-à-dire en les plaçant dans le corpus négatif). L'analyse des vingt-sept CRES restants est particulièrement intéressante et caractéristique des problématiques qui peuvent être rencontrées lors de l'utilisation de ces techniques algorithmiques de TALN.

L'outil a isolé plusieurs types de documents en rapport avec la thématique comme le fait que :

- la contamination corporelle n'est pas avérée mais est indiquée dans le chapitre du CRES relatif aux conséquences potentielles de l'événement [2 CRES], car c'est la totalité du texte qui est analysée par *PLUS* ;
- la contamination corporelle est avérée [15 CRES] : ce sont les documents recherchés ;
- la contamination est avérée mais sur un vêtement, c'est dans ce cas une ambiguïté dans l'écriture à situer dans son contexte [5 CRES] ;
- la contamination corporelle est mentionnée mais sous sa forme négative [5 CRES], il n'y a pas de contamination corporelle. Ces cas ne sont pas considérés comme des faux positifs car il est bien fait mention de contamination corporelle et c'est bien ce que l'on cherche à repérer. La forme négative pouvant amener l'expert à s'interroger plus avant.

On peut conclure de ce travail que l'entraînement d'une dimension est relativement simple et rapide (au regard d'une relecture attentive de tous les CRES indexés), si l'on peut discriminer les exemples positifs et négatifs. Une fois la dimension construite, il reste comme nous l'avons vu précédemment des « erreurs » qui sont principalement dues à des problèmes inhérents à ces techniques algorithmiques, comme la difficulté de prise en compte des formes négatives ou subjectives des phrases. Cependant, cette difficulté est à résoudre en amont de la définition de la dimension. Dans notre exemple, cette difficulté se pose, puisque ce qui est recherché, ce sont les contaminations avérées. Mais cette « intention de recherche » n'est pas explicite et l'algorithme remonte les contaminations suspectées et non avérées, ce qui est légitime puisque les CRES « faux positifs » parlent néanmoins de contamination corporelle.

L'ambigüité narrative peut également amener des faux positifs ou négatifs mais le taux de réussite semble tout de même intéressant et prometteur notamment au regard du temps de construction/entraînement de la dimension (environ 5 hommes/jour, une fois le module disponible dans *PLUS* pris en main et le contenu de la base de CRES relativement maîtrisé) en comparaison de la relecture des 6 200 textes (de 20 pages en moyenne) vis-à-vis de cette thématique ou de recherches croisées multiples et *in fine* très complexes en fonction des sujets traités.

Après avoir isolé les CRES pertinents pour notre problématique des contaminations corporelles, il est possible de les croiser avec d'autres dimensions existantes comme les lignes de défense du modèle A2N-T ou avec d'autres dimensions à créer. On peut par exemple rechercher si les contaminations corporelles sont prépondérantes lorsque l'activité concernée par l'événement fait appel à la « sous-traitance » ou est soumise à une « surcharge de travail », etc. En termes de capacité d'exploration des enseignements à tirer du REX, le champ des possibles est alors quasiment infini et à la discrétion de l'expert.

## IV. RESULTATS ET PERSPECTIVES

L'IRSN, dans le cadre du déploiement de sa feuille de route numérique, mène des réflexions sur « l'expertise augmentée », notamment à travers l'utilisation des technologies digitales pour tirer parti des données collectées ou issues de modèles de simulation. Ces réflexions envisagent

les possibles modifications des pratiques d'expertise par l'utilisation des outils d'IA.

### *A. De premiers résultats encourageants*

Les résultats obtenus par l'utilisation des techniques d'intelligence artificielle et notamment celles de traitement automatique du langage naturel que nous venons de présenter sont prometteurs. Ainsi, la construction des dimensions et leurs utilisations pour identifier automatiquement dans les CRES passés et à venir des défaillances de lignes de défenses et d'autres enseignements variés constituent une voie qui donne des résultats satisfaisants et ouvrent de nombreuses perspectives dans le domaine du traitement du REX. On peut ainsi projeter d'utiliser ces technologies pour réaliser un codage automatique dès la réception des CRES et proposer un usage plus « fluide » du REX dans les activités d'expertise.

Ce premier codage algorithmique permet aux experts et chargés d'évaluation d'identifier rapidement les lignes de défense défaillantes, c'est-à-dire celles ayant un niveau de confiance supérieur à 0,8. Cette première cartographie du CRES ainsi obtenue (à moindre coût) permet d'orienter l'expert vers une lecture immédiate, pour confirmer ou non un questionnement particulier. La construction de dimensions thématiques opportunistes permet également aux experts de se focaliser sur les CRES d'intérêt pour leur domaine de compétence ou leur sujet d'étude. Par exemple, le cas de la « contamination corporelle » traité précédemment peut être d'intérêt pour les experts de la radioprotection qui peuvent de la même façon identifier rapidement une catégorie d'événements pertinents et non repérables à la seule lecture du titre de la déclaration d'événement.

La construction des dimensions est un processus « vivant » a contrario de l'indexation par mots clefs. Elle permet aux experts, au gré de leurs recherches, de leurs envies d'exploration, de l'avancée de leurs sujets d'expertise, de modifier de précédentes dimensions ou d'en créer de nouvelles pour ainsi tester quasiment instantanément l'ensemble de la base disponible. Le passé pourra alors être interrogé sur la base des connaissances actuelles. Contrairement à l'indexation manuelle, la modification d'un élément de recherche ou la création d'un nouvel élément n'implique pas la ré-indexation de l'ensemble des documents, ce qui d'ailleurs n'est jamais réalisé, tant l'énergie à mettre en œuvre est importante.

Avec ce type d'IA, l'expert se dote d'un nouvel outil lui permettant de réinterroger sans cesse la totalité des textes présents dans la base de données. Il s'agit d'une exploration de données (historisées ou à venir) sur la base de critères dynamiques, non figés préalablement par la structure des données. On peut également mettre en exergue la transversalité apportée par ce type d'outil qui permettra de croiser facilement des préoccupations de différents experts. Par exemple, il est assez simple d'extraire des CRES qui font mention de « *contamination corporelle* » associée à une « *charge de travail* » importante. Ces croisements de dimensions construites spécifiquement par domaines d'expertises constitueront les futurs traits d'unions entre disciplines et augmenteront ainsi les connaissances de tous. Ces exemples sont très nombreux et une fois la prise en main faite (accompagnement spécifique, formation des experts à la construction des dimensions et d'une façon plus large aux outils d'IA), il faut compter sur les experts pour croiser, étudier, disséquer, analyser les événements, pour extraire des données, des informations qui produiront des connaissances.

Le dernier point à aborder est celui de l'amélioration continue des dimensions. En effet, leur prise en main par les experts permet une amélioration continue de celles-ci. Chaque garant d'une dimension peut l'affiner pour son compte mais c'est l'ensemble des utilisateurs qui bénéficieront des résultats obtenus en ayant l'assurance que quel que soit le CRES utilisé ils disposeront de la dernière version de la dimension (validée par l'expert du domaine). C'est alors l'ensemble de la communauté qui bénéficie des travaux portés par les experts et donc des connaissances associées. Ces outils mis à la disposition de tous ceux qui voudront bien s'en saisir permettront une aide à la « réflexion » sur les thématiques étudiées.

### *B. L'IA est-elle un expert ?*

Comme l'indique Gaspard Koenig en relatant les propos de Karl Polanyi [6] « *Nous savons davantage que ce que nous pouvons exprimer* », il est évident qu'aujourd'hui, et sur le type d'outil utilisé, la question de l'expert remplacé par une IA n'a pas de sens. David Autor [6] résume cela de la façon suivante : « *l'automatisation sera d'autant plus éloignée qu'une tâche requerra flexibilité, jugement et sens commun* ». L'analyse des CRES rassemble justement ces trois critères, mais aussi la compréhension du contexte, c'est-à-dire du type d'installation, du type d'activité mise en œuvre, de l'organisation en jeu, dans un contexte sociotechnique, industriel et politique particulier. Tous ces éléments non décrits permettent à l'expert de comprendre parfois les sous-entendus, estimer les moyens mis en jeu. Par exemple, l'identification d'une augmentation du nombre de déclarations faisant état d'une contamination corporelle peut être corrélée à des chantiers particuliers en cours ou des opérations de maintenance spécifiques identifiables en dehors de la situation particulière exposée dans un CRES donné. La connaissance contextualisée des installations et des exploitants permet de lire entre les lignes dans un compte rendu officiel à destination d'une autorité. Sans être factuels, ces éléments participent et font l'expertise. Ils sont loin pour l'instant d'être accessibles à l'IA.

L'IA n'est assurément pas un « expert », mais la place que va prendre l'IA dans le rôle de l'expert peut soulever quelques questions et interroger des processus de transition. En effet, si l'utilisation de ces outils d'IA devient la règle, il existe un risque qu'au fil du temps on ne s'interroge plus sur le résultat produit. L'utilisation systématique (« carte blanche ») voire aveugle (en mode « boîte noire ») des résultats proposés par l'IA ne permet plus de s'interroger sur les données initiales, les limites, etc. Cette déviance est déjà observable dans la vie courante. L'IA n'est pas un expert mais devient « *notre expert* » : des algorithmes choisissent les livres ou les films que nous devons aimer, les routes que nous devons emprunter, voire ce qui serait mieux de manger. Les résultats produits par les algorithmes sont pris comme une évidence. C'est l'utilisateur qui lui-même donne à l'IA son statut d'expert et c'est dans ce cas uniquement que ces outils peuvent prendre cette position. Ils passent alors d'un statut d'outil à un statut d'expert. Mais cela n'est pas propre aux outils d'IA : nous étions déjà confrontés à cette problématique avec des « systèmes experts » d'aide à la décision « open source » sur lesquels chacun a amené sa petite « touche personnelle », rendant parfois leurs résultats *in fine* indémontrables.

L'IA ne deviendra « expert » que si l'expert qui l'a formée (entraînée) se réinterroge sur son apprentissage et ses propres pratiques. L'expert devient alors le professeur, il introduit de

fait un biais lié à ses propres limites de connaissances et ses erreurs de jugement. Dans l'exemple présenté, le système devra garder en mémoire que le créateur d'une dimension apporte son biais de perception des faits identifiés dans les CRES. A l'inverse, les utilisateurs se doivent de conserver une attitude critique et interrogative pour ne pas mécaniquement utiliser les résultats de l'IA, validant ainsi les biais et les erreurs d'apprentissage. Ces algorithmes non suivis continueront invariablement à produire des résultats même s'ils dérivent jusqu'à être complètement faux.

Ce qui est évident, c'est que ces outils permettent la réalisation d'un grand nombre de calculs et qu'ils ont la capacité d'analyser, croiser, interroger des quantités de données inaccessibles à l'expert, en tout cas, dans un temps toujours plus contraint. L'IA n'est pas un expert mais devient un outil incontournable de l'expertise.

### *C. Symbiose ou antibiose ?*

« *Le docteur Langlotz prône comme Kasparov dans le domaine des échecs, une étroite collaboration entre l'homme et la machine, qui suppose de former les radiologues à l'IA … et de faire contrôler l'IA par des radiologues.* » [6]. En effet, il nous semble particulièrement intéressant d'envisager l'avenir de ces outils d'IA comme de nouveaux systèmes « experts ». Ces outils sont peut-être les futurs traits d'union entre des domaines de compétence divers et des acteurs variés (experts, généralistes, managers…). A l'image des outils d'IA utilisés aujourd'hui couramment, comme Waze, c'est la communauté qui les construit, les améliore, les enrichit. Ce sont les données et les connaissances que chacun a à sa disposition qui permettent d'améliorer les résultats de ces outils. Les émergences attendues ne seront possibles qu'à condition que l'apprentissage soit de plus en plus fin, que les données utilisées soient de la meilleure qualité possible et que chacun se réinterroge constamment sur les résultats produits. Cela modifiera certainement nos organisations car certains seront garants des données d'entrée, d'autres de l'apprentissage et de son suivi, d'autres encore des résultats produits, etc., et ce n'est qu'à cette condition que nous pourrons dans notre domaine de compétences asseoir notre expertise future et augmentée. Il nous faudra toujours utiliser l'IA comme une extension de notre expertise, un outil facilitant la transversalité et la production de connaissances.

## REFERENCES


[1] Rousseau, J.-M., Montmeat, A., Hebraud, C. et Ayadi, B.-M., « REX et digital : pas de bras, pas de Big Data ! », Congrès de l'IMdR λμ21, Reims, octobre

[2] Evrard, J-M., « Réflexions sur la défense en profondeur pour la sûreté des réacteurs nucléaires », Journée IMdR 'Nouvelles avancées en matière de Défense en profondeur', 12 mars 2020, Cachan.

[3] Rousseau, J.-M., Montmeat, A., Hebraud, C. et Ayadi, B.-M., « Analyse d'événements dans l'industrie nucléaire : la recherche de configurations émergentes comme alternative à l'utilisation de relations causales déterministes », in: J.-F. Vautier (sous la direction de) De la configuration dans des approches systémiques pour appréhender la complexité, Editions Techniques de l'Ingénieur, France, 2018.

[4] Lagarde, S., Raynal, C. et Urieli, A., « Repérer des dimensions dans les REX : utilisation du TAL en milieu médical », Congrès de l'IMdR λμ21, Reims, octobre 2018

[5] Tanguy, L., Tulechki, N., Urieli, A., Hermann, E. & Raynal, C., 2015, « Natural language processing for aviation safety reports: From classification to interactive analysis », *Computers in Industry*, Elsevier.

[6] Koenig Gaspard, «La fin de l'individu », l'observatoire, 2019.