PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à MINES ParisTech

FORAGE DES DONNÉES ET FORMALISATION DES CONNAISSANCES SUR UN ACCIDENT

Le cas *Deepwater Horizon*

Soutenue par

Thibaut EUDE

Le 18 décembre 2018

Ecole doctorale n°432

SCIENCES DES METIERS DE L'INGENIEUR

Spécialité

SCIENCES ET GENIE DES ACTIVITES A RISQUE

Composition du jury :

Gilles MOTET Professeur, INSA Toulouse		*Président*
Aldo GANGEMI Professeur, University of Bologna		*Rapporteur*
Franck GUARNIERI Professeur, CRC Mines ParisTech PSL research university		*Directeur de thèse*
Sébastien TRAVADEL Maître assistant, CRC Mines ParisTech PSL research university		*Maître de thèse*
Jean-Christophe LE COZE Responsable études et recherche, INERIS		*Examinateur*

MINES ParisTech | PSL

Remerciements

Je tiens à adresser mes plus sincères remerciements à Franck Guarnieri et Sébastien Travadel, qui, grâce à leur intelligence de la situation et leur bienveillance, m'ont permis de trouver le chemin à suivre. Ils ont été là, tout le temps, et j'ai grandement apprécié l'esprit d'équipage ainsi formé.

Franck a été le mentor que je recherchais, bien au-delà de la direction de thèse, robuste et en confiance ; honneur m'a été donné de faire partie du *crew* de ce *battleship* qu'est le CRC ; Sébastien a été le cornac, le guide patient et exigeant, qui m'a permis au jour le jour de développer mon esprit scientifique et de me confronter à mes propres limites, ce que je poursuivais depuis longtemps. Et cela ne nous a jamais empêché de bien rire et de profiter de la vie ! J'ai vraiment vécu ma thèse grâce à vous.

Je tiens également à remercier les membres de mon jury de thèse qui ont accepté de bien vouloir relire mes travaux et d'apporter leurs conseils avisés et leur évaluation précieuse. Merci à Gilles Motet, rapporteur, qui a présidé le jury. Merci à Aldo Gangemi, rapporteur, pour qui « la première rencontre » s'est faite au travers de l'ontologie *DOLCE* dont il a été l'un des créateurs ; merci Aldo pour avoir répondu à mon courrier et l'intérêt que vous avez montré à mes recherches ; et merci à Jean-Christophe Lecoze pour l'examen de mes travaux.

D'une certaine manière, je tiens à remercier les scientifiques et fonctionnaires américains, pleinement impliqués lors de l'accident de *Deepwater Horizon*, et avec qui j'ai pu échanger et pour certains rencontrer : un merci particulier à Steven Chu, Paul Hsieh et David Westerholm pour le temps accordé, leur accueil chaleureux et leur accessibilité.

Je tiens à saluer les chercheurs du CRC ; merci notamment à Aurélien Portelli et Didier Delaître pour tous les échanges passionnants, leur gentillesse naturelle et qui m'ont aussi beaucoup aidé, que ce soit pour la préparation de soutenance mais aussi pour parler « d'autre chose » que de la thèse : un bol d'air particulièrement bienvenu ! Un remerciement particulier à Wim Van Wassenhove, le responsable du mastère spécialisé MRI, pour avoir été le coach sportif indispensable dans ma recherche d'un corps (plus) performant. Je garde bon souvenir des courses et autres exercices dans les Bouillides ! Je salue également Emmanuel Garbolino pour ses mots toujours pleins d'esprit lors de nos (trop) brefs échanges. Merci enfin à Justin Larouzée et Eric Rigaud pour leurs conseils lors de la dernière ligne droite avant la défense.

Merci à Sandrine et Myriam pour leur disponibilité et leur support. Merci également à Emmanuel Levrat, technicien informatique du site de Sophia, pour son chaleureux support technique et sa disponibilité.

J'exprime également mon amitié aux autres doctorants et docteurs qui ont partagés avec moi cette aventure un peu bizarre qu'est la thèse de doctorat. Salutations particulières à Aissame, Cécile, Clément, Constance, Dahlia, Diana et Martin, mes collègues de la première heure, soutien et courage pour Ragheb et Pierre qui sont dans le dur.

Je remercie mes parents et beaux-parents pour leur compréhension et patience.

Enfin, j'adresse ici un profond remerciement amoureux à celle qui me soutient depuis toujours et qui a eu le cœur à la fois de me suivre dans cette tempête de cerveau

(une mer calme n’ayant jamais fait un bon marin) et de devenir ma femme alors que je n’avais pas encore défendu mes travaux ; merci à Hanelore, ma belle aimée.

In Memoriam

Les travaux de cette thèse sont dédiés à la mémoire des onze travailleurs de l'*offshore* pétrolier qui ont perdu la vie lors de l'accident de la plateforme de forage *Deepwater Horizon* le soir du 20 avril 2010.

Jason Anderson
Aaron Dale Burkeen
Donald Clark
Stephen Curtis
Gordon Jones
Roy Wyatt Kemp
Karl Dale Kleppinger, Jr.
Blair Manuel
Dewey Revette
Shane Roshto

Hope it will never happen again.

Table des matières

> « *The topic of how to organize and make decisions in a major crisis is fascinating and rarely studied. It is a topic worth investigating* »[1] (Dr. Tom Hunter, ancien directeur des laboratoires nationaux Sandia, membre de l'équipe de gestion de l'accident de *Deepwater Horizon*).

Introduction générale de la thèse

Des volumes considérables de données, parfois contradictoires, sont disponibles à la suite d'un accident industriel, que ce soit sous forme de rapports, de documents techniques ou d'articles scientifiques. Ce constat soulève la question de leur collecte, mais aussi de leur pertinence pour étudier la gestion de crise. La présente recherche vise à explorer des modes de forage de ces quantités de « faits » socialement construits et à proposer une méthode pour structurer la connaissance que l'on peut tirer d'un accident à des fins scientifiques.

Les données issues des rapports officiels d'investigation servent de sources d'informations principales voir exclusives à l'académie, mais aussi à l'industrie et toute partie intéressée à propos de l'accident en question : il y a derrière ces rapports d'enquêtes une forte attente sociétale. En retour, les résultats obtenus par les études scientifiques peuvent servir d'orientations à visée opérationnelle ou politique qui modifient considérablement le paysage réglementaire ou normatif des sociétés. Et la justification du choix des données est aussi importante pour la validité de ces résultats que la qualité du raisonnement qui les produit.

La perte totale le 20 avril 2010 de la plateforme de forage *Deepwater Horizon*, opérée par Transocean pour le compte de BP dans le Golfe du Mexique (à une cinquantaine de mille marins de la côte de l'État américain de Louisiane), sera notre étude de cas tout au long du travail. L'accident s'est manifesté en deux temps : l'explosion et le naufrage de la plateforme, entraînant la mort de onze travailleurs ; puis une catastrophe environnementale, la plus grave de l'histoire des États-Unis. Les quatre millions de barils de pétrole qui ont été rejetés dans l'environnement continuent aujourd'hui d'avoir des répercussions sur les activités humaines dans la région du Golfe. Le coût de la catastrophe supportée par BP est estimé à soixante-deux milliards de dollars. L'organisation de la gestion de crise a dû faire face à de nombreux défis dans un contexte d'incertitude et sous une pression intense et permanente générée notamment par les attentes de la société civile et la très grande méconnaissance du phénomène sous-marin. Des solutions d'ingénierie ont été inventées sur place et déployées pour tenter d'abord d'atténuer les dommages à l'environnement puis de sceller le puits le 15 juillet pour enfin le « tuer » le 19 septembre 2010. Reprendre le contrôle du puits devenu sauvage et stopper la pollution aura nécessité quatre-vingt-sept jours et il aura fallu au total cinq mois d'un effort sans commune mesure fournit par la plus grande organisation de gestion de crise jamais formée pour faire face à cet accident.

1 Extrait d'une correspondance personnelle avec l'intéressé.

Nous avons recensé au moins une cinquantaine de documents, dont (presque) une trentaine de rapports d'enquêtes qui traitent de la problématique technique de l'accident. A cela s'ajoutent plus de quatre mille documents écrits, dont les preuves exhibées au procès, le plus lourd contentieux de l'histoire juridique des Etats-Unis à ce jour. Un tel constat doit interpeller le chercheur. Comment fouiller ces masses d'informations ? Comment les organiser pour conduire une étude des mécanismes de gestion de crise ? Les « *safety studies* », ces sciences qui s'intéressent à notre rapport au danger et aux moyens de le régler, ne semblent pas pour autant prendre en considération l'impact de cette évolution.

Ces interrogations renvoient à des problématiques d'extraction de notions telles que la causalité à partir d'une narration, ou à la création de bases de connaissances traçant les sources et les éventuelles contradictions entre les informations. En pratique, la compréhension humaine d'un évènement est sérieusement entravée si elle ne peut s'appuyer sur des outils informatiques de traitement de données. Typiquement, la mise en lumière des mécanismes causaux tissés par les nombreux rapports d'investigation publiés à la suite d'un accident, et destinés à justifier des actions correctrices en réponse à la catastrophe, pourrait fournir un moyen d'appréciation de la robustesse d'un argumentaire, ou pourrait permettre de cerner des zones d'ombre sur lesquels devraient porter les travaux scientifiques.

Notre objectif est à la fois de poser des bases robustes pour aborder scientifiquement ces questions, à travers des hypothèses claires et des choix explicites, et de proposer de premières méthodes de traitement destinées à être automatisées. Le développement de ces méthodes a suivi l'évolution de notre compréhension du cas *Deepwater Horizon* pendant la période de notre travail de thèse. Il s'agit donc de rendre compte de manière formelle de processus au départ informés par une démarche experte. L'exploration experte et systématique de la catastrophe a en effet montré la richesse des informations disponibles, non pas en l'état, mais dès lors que les données étaient confrontées entre elles, les conclusions croisées entre les différents rapports d'investigation, etc. Nous avons ainsi pu mettre en évidence des séquences inexpliquées et pourtant cruciales au cours du processus de gestion de la « marée noire ».

Les résultats de notre travail de recherche sont regroupés en quatre chapitres.

Le chapitre 1 présente la problématique du forage des données produites à la suite d'un accident (1). Afin d'introduire des concepts clés pour la suite, nous réinterrogeons la définition d'« accident » par le biais d'une analyse sémantique (1.1) : nous explorons le noyau de sens du mot d'abord (1.1.1), puis la synonymie étendue de cette unité lexicale (1.1.2) et, enfin, nous proposons un panorama des espaces sémantiques du mot « accident » à travers ses usages, pour en proposer une définition de cadrage de nos travaux (1.1.3). De là, nous nous penchons sur les données produites à la suite d'un accident, à travers l'exemple des documents publiés sur le cas *Deepwater Horizon* (1.2) : nous explicitons notre stratégie de collecte (1.2.1), et la problématique de qualité et d'accessibilité des données *via* le Web (1.2.2), avant de porter notre attention sur les rapports d'enquête (1.2.3). Ensuite, nous présentons deux travaux de forage (1.3) : d'abord une démonstration de la lacune flagrante du traitement de ce cas

d'accident par la science (1.3.1.), puis une présentation détaillée de la construction d'une donnée critique, une métrique de l'accident (1.3.2), exemple qui rend évident la nécessité d'évaluer et d'organiser les connaissances à propos d'un accident à des fins de science (1.3.3).

Le chapitre 2 montre la pertinence de recourir au concept d'ontologie en tant que produit de l'ingénierie des connaissances pour résoudre notre triple problème de recensement des données, de leur mise en relation conceptuelle et de leur représentation (2). Nous présentons ce qu'est une ontologie sur le plan théorique (2.1), depuis son origine philosophique à son arrivée en science (2.1.1). Puis nous présentons les travaux de « grands ontologues » qui ont pensé l'ontologie logique et informatique (2.1.2) et rappelons les avantages des ontologies comme outils de traitement de la connaissance (2.1.3). Nous présentons par la suite, l'ingénierie, la construction et les opérations des ontologies (2.2), soit : les grands principes de logique formelle pour leur construction (2.2.1), les opérations rendues possibles sur les connaissances (2.2.2) et enfin, leur implémentation informatique à partir d'un langage spécifique (2.2.3). Enfin, nous montrons comment nous en sommes venus à choisir l'ontologie de haut niveau *DOLCE* pour résoudre notre problème (2.3) : après un tour d'horizon de quelques réalisations informatiques construites à partir des ontologies (2.3.1), nous présentons en détail notre cheminement de pensée pour partir à la recherche d'une ontologie « idéale » (2.3.2), jusque à *DOLCE*, qui s'est imposée à nous par le décryptage des fondamentaux de sa conception (2.3.3).

Le chapitre 3 présente notre proposition d'algorithmes de peuplement d'ontologies, élaborée à partir des observations exposées dans le chapitre 1 et des fondations théoriques montrées au chapitre 2. Nous proposons une théorie et une méthode pour à la fois peupler une ontologie de façon automatique avec des instances d'évènements et être capable de traiter la causalité exprimée dans un texte (3). Notre proposition se situe dans le domaine du Traitement Automatique du Langage Naturel (TALN). Nous exposons d'abord le processus de population d'une ontologie (3.1) : après avoir précisé les concepts linguistiques de morphologie et de syntaxe (3.1.1), la sémantique et le processus de lemmatisation (3.1.2), nous présentons notre solution, soit un algorithme *Named Entity Recognizer* susceptible de procéder à la création d'instances dans une ontologie telle que *DOLCE* (3.1.3). Ensuite, nous exposons notre méthode pour le traitement automatique de la causalité exprimée dans un texte (3.2) : après un rappel théorique du raisonnement causal et de son expression dans un texte (3.2.1), et un bilan des apports récents de l'Intelligence Artificielle (IA) dans la branche *Question Answering* (3.2.2), nous proposons une solution algorithmique, construite à partir d'une méthode probabiliste bayésienne, travaillant à la fois sur la sémantique et la syntaxe, renforcée par la mise en réseau des classifieurs (3.2.3). Enfin, nous présentons l'architecture de ce que sera notre machine qui répond à la question « pourquoi ? » (3.3) : nous exposons en premier ce que nous entendons par cheminement causal (3.3.1), puis nous amenons une preuve de concept avec un essai simulé sur un texte de notre machine (3.3.2) ; enfin, nous clôturons ce chapitre par un embryon de cahier des charges pour son implémentation informatique et par des projections sur des utilisations possibles de notre machine (3.3.3).

Dans le chapitre 4, nous présentons nos résultats issus de l'étude du cas *Deepwater Horizon* (4). Nous montrons une ontologie de l'accident de *Deepwater Horizon* (4.1). Nous procédons d'abord par une présentation en langage naturel de l'accident (4.1.1). Puis nous présentons notre ontologie du cas (4.1.2). De là, nous montrons quelques cas concrets d'utilisations possibles de notre ontologie et de sa population (4.1.3). Ensuite, nous allons montrer comment une ontologie peut aider à orienter des travaux de recherche (4.2). Nous allons présenter comment une représentation graphique ontologique nous a permis de mettre en lumière un « incident d'explication » à propos d'une séquence de l'intervention (4.2.1). Puis, suite à cette découverte, nous montrons notre travail de recherche de l'explication manquante (4.2.2). Enfin, émergeant de cette explication, nous montrons que c'est un raisonnement contrefactuel qui est à l'origine de la prise de décision la plus critique de l'intervention (4.2.3). Enfin, la dernière partie de ce chapitre est consacrée à la discussion de notre proposition énoncée au chapitre 3 et des résultats présentés dans ce chapitre (4.3). Nous ferons le tour des limites des ontologies (4.3.1), des limites de notre machine IA, notre apport présenté au chapitre 3 (4.3.2) et enfin des limites générales à notre étude du cas *Deepwater Horizon*.

Chapitre 1 L'« accident » et la problématique du forage des données

Dans ce chapitre, nous introduisons la problématique du forage des données sur un accident. Dans un premier temps, l'interrogation du sens attribué au terme « accident » lui-même, à travers son usage dans la langue, permet de cerner notre objet d'étude, et de mieux repérer les différentes perspectives adoptées pour étudier un événement ; cette longue première section vise également à introduire des concepts-clés de sémantique mobilisés dans les chapitres suivants (1.1). Nous aborderons d'abord l'accident par une analyse qualitative de son espace sémantique (1.1.1), puis par une analyse de ses synonymes (1.1.2) et enfin nous proposerons un ensemble de représentations de l'espace sémantique de l'« accident » pour en tirer une définition nouvelle. Dans un second temps, nous présentons les données relatives au cas *Deepwater Horizon* et soulignons leurs caractères problématiques dans un contexte d'étude scientifique (1.2). Nous présentons d'abord notre stratégie de collecte de l'information pour constituer un corpus documentaire pertinent (1.2.1). Ensuite, nous présentons les abords de l'accident et sa consistance en matière de données accessibles *via* le Web et notre approche de l'évaluation de la donnée (1.2.2). Enfin, nous présentons un panorama et quelques statistiques descriptives à propos des rapports d'enquêtes élaborés suite à cet accident, matériau de base choisi pour notre recherche (1.2.3). Enfin, nous présentons un premier forage de données effectué à l'aide de méthodes d'analyses bibliographiques des « connaissances » disponibles sur ce cas, soient les conclusions scientifiques qui en ont été tirées à ce jour (1.3). Nous montrons d'abord la lacune de traitement à propos de ce cas de la science et dans ses conclusions (1.3.1), puis nous illustrons l'aspect de construit social de la donnée par le biais d'une analyse fine d'un attribut quantitatif lié à l'accident, « *oil budget* » (1.3.2). Enfin, après cet ensemble de démonstrations, nous concluons sur l'impérieuse nécessité d'organiser les données aux fins d'études de science (1.3.3).

1.1 Le concept d'accident : analyse sémantique

L'unité lexicale « accident » est polysémique et est associée à divers concepts en fonction de ses usages. Dans cette section, nous recherchons les principales significations véhiculées par cette unité lexicale. Nous partons d'une analyse qualitative du sens commun du mot puis nous analysons son espace sémantique[2] selon la méthode exposée par les concepteurs du dictionnaire électronique des synonymes (DES) élaboré par le CRISCO[3] et exposée par Manguin et al. (2004).

1.1.1 Analyse qualitative de l'espace sémantique

Nous nous appuyons sur le modèle sémantique proposé par Victorri et Fuchs (1996a). A une unité lexicale polysémique sont associés deux espaces continus : l'espace sémantique dans lequel le sens d'une expression dans un énoncé est représenté par une région et l'espace co-textuel qui représente la détermination du sens d'une expression par les autres éléments présents dans l'énoncé. Victorri et Fuchs (1996b) modélisent alors la « dynamique » de construction du sens par une fonction de l'espace co-textuel sur l'espace sémantique qui, à tout élément de l'espace co-textuel, associe une fonction de potentiel déterminant la dynamique sur l'espace sémantique (i.e. au minimum de cette fonction sont associés des bassins d'attracteurs de sens). En pratique, dans une première étape, nous recherchons à partir des définitions de sens commun du mot « accident » la structure de son espace sémantique.

1.1.1.1 Le noyau de sens

Pour la suite, toutes les définitions sont celles proposées par le CNRTL[4] et les références des définitions correspondent à celles de ce corpus. Nous retenons les définitions suivantes du mot « accident » :

- « *Ce qui s'oppose à la substance ou à l'essence* » *(définition I.) ;*
- « *[P. oppos. à substance]. Ce qui existe, non en soi-même, mais dans un autre ; (…) par ex., la couleur, la forme, qui ne peuvent être que la couleur, ou la forme de quelque chose subsistant en elle-même* » *(définition I.A.1.) ;*
- « *[P. oppos. à essence]. Ce qui ne fait pas partie de la nature ou de l'essence d'un être et peut devenir autre sans qu'il y ait changement d'espèce. Par exemple le fait d'être assis, ou couché, d'être à Paris, d'être en face de Pierre* » (définition I.A.2.) ;
- « *Dans ces emplois le sens glisse insensiblement vers l'idée générale de variation, de variété, qui pour l'œil de l'observateur rompent la monotonie du fond* » (définition I.C) ;
- « *Évènement fortuit, sans motif apparent et sans lendemain, qui affecte une personne ou un groupe de personnes, en interrompant le déroulement normal, probable et attendu des choses* » (définition II).

2 i.e. les mots associés dans les usages de synonymie.
3 Source : http://crisco.unicaen.fr/des/
4 Source : http://www.cnrtl.fr/

Nous retenons comme valeur primaire le sens de la définition I.C, à savoir l'idée générale de variation qui pour l'œil de l'observateur rompt la monotonie du fond. Les autres définitions paraissent resserrées autour de cette définition primaire, dont on peut déduire un « noyau de sens », c'est-à-dire une description minimale valable pour toutes les valeurs de « accident » (Victorri and Fuchs, 1996, p. 66). Ce noyau de sens peut être exprimé à l'aide d'un schéma (Travadel and Guarnieri, 2018). On se donne un domaine *D* quelconque (temporel, spatial, notionnel, etc.) et une proposition *P* dont le domaine de validité *D(P)* est une partie de *D*. Enfin on se donne une trajectoire *T* à travers *D* et un instant t_0 :

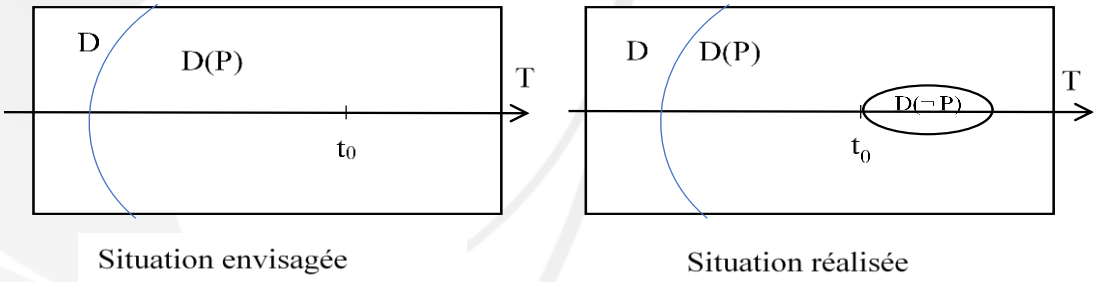


Figure 1 : noyau du sens « accident »

Les définitions précédentes de « accident » ont en commun de souligner qu'à l'instant t_0 on rencontre une frontière du domaine de validité *D(P)* alors qu'il était envisagé son extension au-delà de t_0.

Par exemple, si le domaine *D* est interprété comme une extension spatiale, le domaine *D(P)* comme l'espace de validité d'une propriété d'un objet et la trajectoire *T* en tant que description dans le temps (parcours de l'espace) de cet objet, alors l'accident correspond à la découverte d'une propriété qui se détache de ce domaine spatial en ce qu'elle contredit la propriété *P* (définition I.A.1.). Si le domaine *D* est interprété comme la notion d'être humain et *P* une posture, la trajectoire *T* à travers *D(P)* est susceptible de se prolonger au-delà de t_0 ou peut soudainement rencontrer une frontière de *D(P)* sans que ceci ne remette en cause l'existence de *D* (définition I.A.2.).

La définition II, à savoir « l'accident » comme « *évènement fortuit (…) interrompant le déroulement normal (…) des choses* », renvoie à l'idée de fin de validité de *P* au-delà de t_0 le long de la trajectoire *T* :

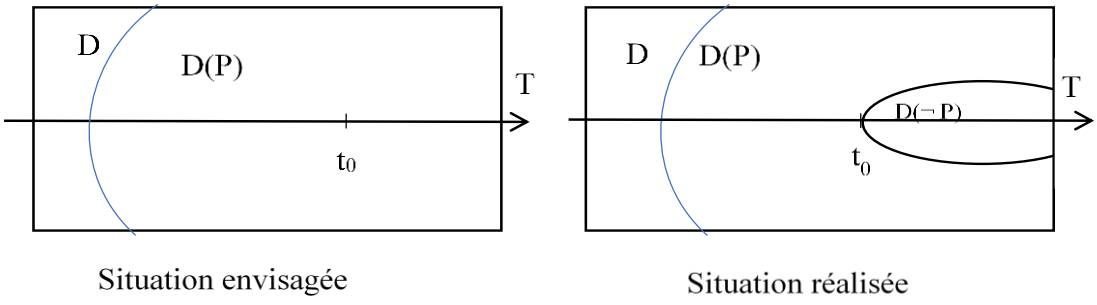


Figure 2 : interprétation schématique de la définition II

Cette définition de sens commun nous apparaît comme la plus en lien avec notre sujet d'étude, à savoir les accidents industriels. Nous en précisons les termes dans la suite du document.

Ici « l'évènement » désigne : « *tout ce qui se produit, tout fait qui s'insère dans la durée* » (définition B.1.). « Normal » renvoie à ce qui est prévisible, conforme à une règle, une norme, un standard et attendu (habituel, compréhensible). Nous retenons pour la définition de la normalité ce « *qui est dans le cours ordinaire des choses, qui se déroule selon un processus considéré comme courant* » (définition 3.a.). L'accident semble alors une compromission des critères de normalité ; nous pouvons rapprocher cette acception d'une forme de perte d'équilibre ou de stabilité, telle que proposée par Wilde (1982) dans son modèle homéostatique des systèmes. « Fortuit » désigne ce « *qui se produit par hasard, de manière imprévue* ». « Motif » a pour synonymes premiers « cause » et « raison » ; « apparent » désigne ce qui est « *visible, perceptible au regard ou à l'entendement* » ; on en déduit qu'est « sans motif apparent » ce qui échappe à la compréhension et qui par conséquent peut paraître fortuit et contraire à la normalité. Quant à « sans lendemain », l'expression renvoie à un caractère « *éphémère* ». La définition de l'accident exprime des effets sensibles, non pas par lui-même (dans sa dimension phénoménologique), mais de par l'interruption du cours normal de la vie qu'il engendre. Si, *a priori*, la cause de l'accident peut être imperceptible, ce sont ses effets qui lui donnent à exister au travers de ce qu'il a perturbé. En d'autres termes, un accident sans cause peut exister, mais pas un accident sans effet. Ce n'est que pris comme objet d'étude que l'accident peut révéler sa nature, son essence, ses raisons.

1.1.1.2 Environnement sémantique direct et co-textualité de l'unité lexicale « accident »

Dans un second temps, nous examinons comment l'unité lexicale « accident » est utilisée dans la langue et environnée dans le champ lexical, afin de cerner d'éventuelles déterminations co-textuelles des différentes valeurs de l'espace sémantique. En effet, avec Victorri et Fuchs (1996, chap. 3) nous supposons que l'espace sémantique d'un mot (i.e. l'ensemble des sens que peut revêtir ce mot) est structuré par d'autres mots auxquels il est associé dans la langue. Les associations à d'autres mots déterminent ainsi des ensembles de significations que peut prendre l'unité lexicale ; l'unité lexicale en question possède des « valeurs typiques » (les centres des bassins d'attracteurs dans l'espace sémantique considéré). Ces valeurs typiques ont la particularité d'être stables et de livrer un cadre interprétatif invariant pour l'unité lexicale en question. Victorri et Fuchs précisent que les valeurs typiques sont des « dérivations » de la valeur primaire de l'unité lexicale en question et que c'est « *[…] l'analyse de ces glissements qui permet de définir correctement le noyau de sens, comme justement la partie du sens qui reste invariante lors de ces modifications* » (Victorri and Fuchs, 1996, pt. 5.3). Pour cela, nous procédons à une analyse du co-texte, c'est-à-dire de l'environnement lexical de l'unité « accident », formé des syntagmes[5] minimaux jusqu'à la phrase (notre limite pour l'étude). Il s'agit de comprendre à

5 Nous explorons ce concept plus en détail dans le chapitre 3.

travers les usages du mot s'il existe une induction de l'espace co-textuel vers l'espace sémantique génératrice de nouveaux sens, c'est-à-dire des valeurs typiques en sus du noyau de sens.

Nous suivons la méthode proposée par Victorri et Fuchs (1996, chap. 6). Il faut en premier lieu recenser tous les « énoncés occurrences » où l'unité lexicale « accident » est utilisée ; c'est virtuellement impossible. Aussi, nous nous contentons de corpus de textes que nous supposons représentatifs de l'utilisation de la langue française dans toute sa richesse et diversité. Pour cela, nous avons choisi le corpus proposé par *Sketch Engine* (« Sketch Engine | language corpus management and query system, » n.d.) qui est le corpus *frTenTen12* (« French Web corpus (frTenTen) search | Sketch Engine, » n.d.; Jakubíček et al., 2013 ; Suchomel and Pomikálek, 2012). Le corpus *frTenTen12* contient près de dix milliards de mots (« Corpus info : French Web 2012 (frTenTen12), » n.d.). *Sketch Engine* permet une exploration en profondeur de larges quantités de corpus, où l'on peut identifier des combinaisons de mots d'intérêt, dans notre cas des co-occurrences (« Concordance | Sketch Engine, » n.d.), c'est-à-dire des mots qui apparaissent dans la proximité syntaxique immédiate (que nous assimilerons au co-texte) d'une unité lexicale ciblée.

L'analyse porte sur toutes les formes issues de l'unité lexicale « accident », le lemme de base[6] (accidenter, accidentel, etc.)[7]. Nous procéderons par recoupements successifs, c'est-à-dire en utilisant différentes approches d'analyse de corpus pour faire émerger la co-textualité et son éventuelle influence.

1.1.1.2.1 Co-occurrences

On recherche d'abord les co-occurrences les plus « fortes », selon l'algorithme *logDice* (Rychlỳ, 2008 ; « Statistics used in Sketch Engine | Sketch Engine, » n.d.), entre « accident » et les autres éléments du corpus. L'algorithme *logDice* est basé sur l'indice de Dice qui s'exprime de la sorte :

$s = \frac{2|X \cap Y|}{|X|+|Y|}$, avec X et Y deux éléments que l'on souhaite comparer (équation 1).

« Accident » apparaît 591 626 fois dans le corpus (en novembre 2017) ; il est précisé que l'analyse des co-occurrences couvre 63,9 % de l'ensemble des occurrences d'« accident ». Nous présentons dans la suite quelques extraits des co-occurrences établies pour chaque catégorie grammaticale. Dans chaque tableau, la seconde colonne indique le nombre de co-occurrences et la troisième le classement selon la valeur donnée par l'algorithme *logDice*.

6 Un « lemme » est une « forme graphique choisie conventionnellement comme adresse dans un lexique » (Ploux and Victorri, 1998a) Nous revenons sur ce concept en profondeur dans le chapitre 3.
7 L'ensemble des données est en annexe.

Premier tableau :

Adjectif	Fréquence	score logDice
« **modifier** »	90785	15,34
vasculaire	13986	11,69
mortel	9294	10,46
grave	6345	8,75
nucléaire	5040	8,64
domestique	3152	8,63
ischémique	927	8,31
corporel	2618	8,3

Tableau 1 : adjectifs associés à « accident »

Les adjectifs sont descriptifs ou amènent une relativité (« grave ») et précisent le « domaine » de la « monotonie du fond » ; ils n'altèrent pas la valeur primaire « d'accident ».

Second tableau :

Conjonction de coordination	Fréquence	score logDice
et_ou	108287	18,3
Incident	3469	8,45
Maladie	11868	8,11
Infarctus	501	6,98
Myocarde	430	6,84
Incendie	1001	6,72
Suicide	894	6,72
Blessure	1500	6,64

Tableau 2 : termes associés à « accident » avec le « prédicat » de conjonction « et_ou »

Ici, ce sont les unités lexicales utilisées en conjonction avec « accident ». « Maladie », « infarctus », « suicide », « incendie » et « blessure » apportent une précision concernant la « rupture de monotonie », en la connotant de façon négative, notamment relativement aux effets sur le corps. Il s'agit d'un possible enrichissement sémantique, à condition de savoir s'il émerge une valeur typique propre à cette co-textualité. « Incident » est l'unité lexicale la mieux classée et la deuxième plus utilisée en conjonction avec « accident ». Elle apparaît 229 153 fois dans le corpus. Attardons-nous sur son cas.

1.1.1.2.2 Word sketch difference

Nous allons utiliser le moteur *word sketch difference* (« Word Sketch Difference – compare collocations | Sketch Engine, » n.d.) pour étudier deux lemmes : nous regardons comment « accident » et « incident » sont articulés à l'aide de l'adjectif qualificatif car « *[…] il ajoute à ces substantifs l'idée des qualités ou des manières d'être sous lesquelles ils sont considérés* » (Ploux and Victorri, 1998). On regarde donc

les deux utilisations en grammaire française de l'adjectif, à savoir comme épithète (prédicat : *modifier*) et comme attribut (prédicat : *adj_sujet_of*).

L'analyse de milliers de concordances où les lemmes « accident » d'un côté et « incident » de l'autre sont utilisés permet de voir si des « *patterns* » d'utilisation se dégagent. On obtient les résultats suivants :

modifier	90,785	34,088	0.15	0.15
vasculaire	13,986	27	11.7	3.4
corporel	2,618	15	8.3	1.2
mortel	9,294	124	10.5	4.7
ischémique	927	7	8.3	2.6
cérébrovasculaires	323	3	6.8	1.5
cérébral	827	16	7.2	2.0
domestique	3,152	105	8.6	4.1
cardio-vasculaire	906	31	8.0	4.0
cardiaque	1,457	122	7.5	4.2
nucléaire	5,040	517	8.6	5.6
cardiovasculaires	602	41	7.4	4.5
grave	6,345	1,970	8.7	7.2
tragique	977	316	7.4	6.3
majeur	2,272	2,073	6.8	6.8
imprévu	223	231	5.8	6.6
malheureux	370	559	6.2	7.4
mineur	238	809	5.1	7.3
fâcheux	161	402	5.6	8.1
regrettable	139	398	5.4	8.1
signalé	32	104	3.4	6.4
isolé	74	455	3.9	7.2
violent	99	1,191	3.2	7.0
notable	54	525	3.5	7.4
diplomatique	51	2,141	2.8	8.6
frontalier	0	238	--	6.6

adj_sujet_of	3,855	1,626	0.01	0.01
arrivé	16	0	4.8	--
négligeable	14	0	3.9	--
infinitésimal	3	0	3.6	--
imprévisible	30	3	4.4	1.1
probable	35	6	4.0	1.5
inévitable	111	25	5.9	3.8
fréquent	293	71	6.2	4.2
imputable	74	18	7.0	5.1
prévisible	44	12	4.7	2.9
rarissime	22	6	5.9	4.3
évitable	31	10	6.6	5.3
rare	298	130	4.9	3.7
minime	18	8	3.6	2.5
attribuable	44	19	6.0	5.0
grave	170	101	3.8	3.1
bénin	10	8	5.1	5.1
gravissime	3	5	3.5	4.7
regrettable	11	30	4.3	5.9
révélateur	9	53	2.9	5.5
intentionnel	0	7	--	3.6
prétexte	0	3	--	3.7
condamnable	0	4	--	3.7
symptomatique	0	8	--	3.7
fâcheux	0	10	--	4.3
facheux	0	3	--	5.6

Tableau 3 : Word sketch difference, adjectifs associés à « accident » et « incident », utilisés comme épithète ou comme attribut

Le code couleur est le suivant :

Tableau 4 : code couleur de la distribution des utilisations entre « accident » et « incident »

Le calcul de la différence permet d'illustrer l'utilisation d'un élément par rapport aux deux lemmes comparés et de voir, par exemple, s'il existe des associations typiques (avec un lemme ou un autre) ou génériques (les deux).

On en déduit qu'« accident » est nettement associé aux atteintes à l'intégrité physique, et « incident » à la portée immorale d'un évènement.

En revanche, il semble qu'il n'y ait pas de démarcation flagrante entre « incident » et « accident » sur les dimensions de la fréquence et, plus surprenant, de la gravité en tant que telle (sauf à considérer une telle distinction sur le seul plan de l'atteinte à l'intégrité physique). En effet, nous pouvons regrouper les adjectifs les plus génériques comme tel :

Gravité	Fréquence	Explication
grave	prévisible	imputable
tragique	évitable	attribuable
majeur	rare	
minime	rarissime	
bénin	imprévu	
gravissime		
malheureux		
regrettable		

Tableau 5 : adjectifs associés en commun avec « accident » et « incident »

Plus précisément, on s'aperçoit que si l'on souhaite évoquer l'une des deux dimensions citées précédemment (« gravité » et « fréquence ») à propos d'un « accident » ou d'un « incident », il semble qu'il faille nécessairement y associer un des adjectifs vus ci-dessus. L'examen des concordances où apparaissent les co-occurrences « accident » et « majeur », puis « incident » et « majeur », confirme qu'il est difficile d'y trouver des nuances significatives qui permettent de distinguer dans l'usage « accident » ou « incident » sur les dimensions que nous avons inspectées. Un « incident » n'est pas un « accident » en plus ou moins grave. De là, on peut aussi dire que l'espace co-textuel ne génère pas non plus d'induction dans l'espace sémantique sur ces dimensions en question ; le noyau de sens n'est pas élargi. On observe le même phénomène avec l'adjectif « rare ».

Toutefois, une approche *Longest-Commonest Match* (*LCM*)[8] montre que si de nombreux « *LCM* » construits dans le domaine médical rappellent l'articulation entre « accident » et « maladie », un « *LCM* » associe « accidents bons ou mauvais », ce qui élargit le sens d'accident vers une connotation « positive » et non exclusivement négative. Nous tâcherons dans la suite de montrer si cela structure de manière significative l'espace sémantique.

1.1.1.2.3 Thesaurus

Nous continuons notre exploration de l'espace co-textuel avec la suite logicielle *Sketch Engine* et l'utilisation d'un thésaurus généré spécifiquement autour du lemme « accident ». L'algorithme d'élaboration du thésaurus s'appuie sur le calcul d'une distance entre deux lemmes (« Statistics used in Sketch Engine | Sketch Engine, » n.d.). L'indice *logDice* fournit une liste de co-occurrences. Ensuite, une analyse grammaticale automatique des corpus détecte lorsque l'unité lexicale « nœud » est remplacée dans la phrase par une autre identifiée comme « pertinent » selon l'indice *logDice*. Par exemple, lorsqu'un même texte contient les phrases « l'accident a fait de nombreux morts » et « la catastrophe a fait de nombreux morts », l'algorithme considèrera « accident » et « catastrophe » similaires, sachant qu'ils sont fortement associés par l'indice *logDice*. Cette méthode repose sur l'hypothèse de sémantique distribuée (Schütze and Pedersen, 1995).

8 Le *longest-commonest match* (*LCM*) a été développé par Kilgarriff ('Adam Kilgarriff bibliography | Sketch Engine', no date; 2015) Il s'agit de déterminer quelle est la réalisation syntaxique la plus fréquente et la plus « longue » (en nombre de mots) contenant une co-occurrence.

Le thésaurus généré à partir du corpus *frTenTen12* autour du lemme « accident » contient soixante mots. Là encore, les notions portées par les lemmes de ce classement semblent déjà contenues dans le noyau de sens d'« accident ». On trouve les lemmes les plus significatifs suivants :

Catastrophe	0,367	Crise	0,292	Conflit	0,248
risque	0,33	circonstance	0,282	rupture	0,247
conséquence	0,316	Dégât	0,269	fuite	0,245
perte	0,311	changement	0,265	crime	0,245
dommage	0,307	Situation	0,256	état	0,244
difficulté	0,306	Problème	0,255	phénomène	0,243
évènement	0,303	Condition	0,253	acte	0,243
chute	0,296	Impact	0,252	cause	0,238
danger	0,295	Retard	0,25		
erreur	0,294	Fait	0,25		

Tableau 6 : items de la classe « accident »

Cela confirme que la valeur primaire d'« accident » est porteuse d'un jugement de valeur, très majoritairement négatif.

Nous représentons la variation de l'espace sémantique déterminée par l'induction co-textuelles selon une dimension (Victorri and Fuchs, 1996a, pt. 3.4), nommée « jugement de valeur », et génératrice de deux valeurs potentielles du sens attribué à « accident ».

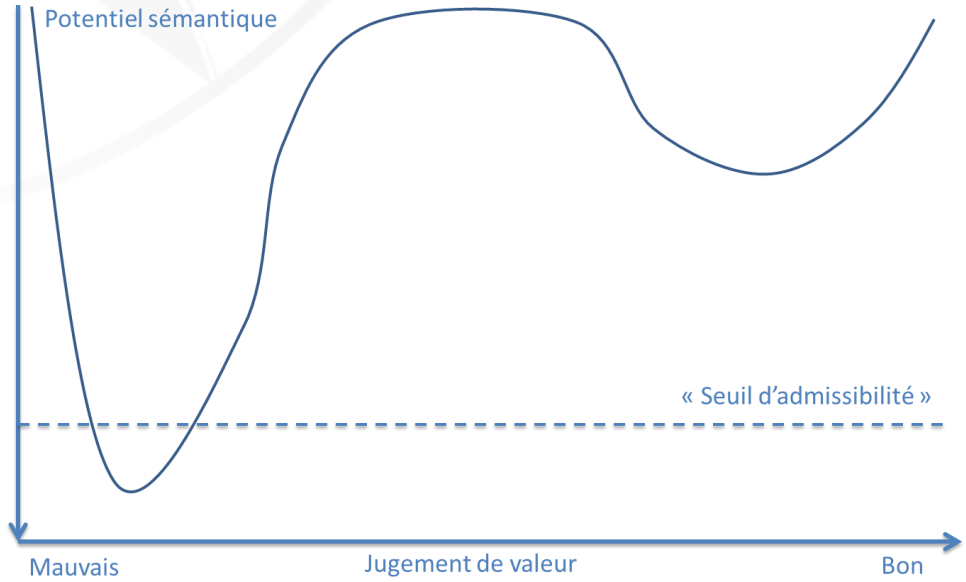


Figure 3 : l'espace co-textuel induisant la connotation de jugement de valeur dans l'espace sémantique d'« accident »

La figure 3 montre alors que l'on peut considérer que le potentiel sémantique U_p déterminé par le paramètre « jugement de valeur » franchit le « seuil d'admissibilité » pour déterminer l'espace sémantique d'« accident » dans sa connotation maligne ; mais la connotation positive est toujours associée à un adjectif qualificatif sans ambiguïté sur sa valeur primaire (« bon », « heureux »). De plus, d'un point de vue

statistique, ces dernières co-occurrences sont particulièrement faibles dans le corpus considéré (mille trente-quatre et cent-vingt-cinq co-occurrences respectivement).

A l'issue de cette brève analyse, nous considérons qu'il n'y a pas de « catastrophe de sens » à l'intérieur de l'espace sémantique d'« accident », sa structure est constante, centrée sur la connotation négative. Pour mémoire, à l'issue de l'examen qualitatif, nous avions retenu la définition primaire suivante : « *évènement fortuit, sans motif apparent et sans lendemain, qui affecte une personne ou un groupe de personnes, en interrompant le déroulement normal, probable et attendu des choses* » (définition II). Nous avons montré ici que le mot ne possède pas dans son noyau de sens, de dimension de gravité, de fréquence ou d'explication. Il faut pour cela utiliser un co-texte qui amènera à l'« accident » la dimension souhaitée. Par ailleurs, ce qui « affecte » une personne ou un groupe de personnes correspond, dans l'usage du mot, à un dommage physique ou matériel, notamment une atteinte à l'intégrité physique.

Dans la suite, nous préciserons la structure de l'espace sémantique du « mot » à partir de ses synonymes.

1.1.2 Analyse synonymique de l'accident : aspect théorique

La langue française est marquée d'une forte polysémie, de sorte que le sens nouveau « *quel qu'il soit, ne met pas fin à l'ancien. Ils existent tous les deux l'un à côté de l'autre. Le même terme peut s'employer tour à tour au sens propre ou au sens métaphorique, au sens restreint ou au sens étendu, au sens abstrait ou au sens concret… A mesure qu'une signification nouvelle est donnée au mot, il a l'air de se multiplier et de produire des exemplaires nouveaux, semblables de forme, mais différents de valeur* » (Bréal cité dans Victorri and Fuchs, 1996a, p. 4).

Partant de ce constat, nous cherchons à préciser la signification de l'unité lexicale « accident », à travers les « significations » qu'il véhicule dans ses usages. Le « sens » d'un énoncé est entendu ici comme « *la contribution constante du matériau linguistique… au sens de toute occurrence de cet énoncé. Autrement dit, nous supposons que [l'énoncé] possède une qualité intrinsèque, qui ne dépend que de sa forme, qui explique sa capacité à produire, dans un contexte donné, un sens pour l'occurrence correspondante de l'énoncé en question. Et c'est cette qualité intrinsèque que nous appelons le sens de [l'énoncé]. Ce sens peut donc être vu comme un "potentiel" de sens en contexte, ou encore comme un ensemble de contraintes linguistiques, qu'il faut ajouter aux conditions d'énonciation pour comprendre la fonction et les effets [de l'occurrence d'un énoncé]* » (Victorri and Fuchs, 1996b, p. 13).

En pratique, nous utilisons le dictionnaire des synonymes proposé par le CRISCO (Manguin *et al.*, 2004), accessible aux non-linguistes pour représenter les relations synonymiques et l'espace sémantique. Nous nous appuyons sur les travaux de Ploux et Victorri (1998b) et une modélisation continue de la polysémie : à chaque unité polysémique est associé « *un espace sémantique, de petite dimension, muni d'une structure mathématique précise, et l'on représente le sens de l'unité dans chacun de ces emplois par une région de cet espace sémantique* » (Ploux and Victorri, 1998a, p. 5). Ce sont les agencements spatiaux des régions et sous-régions de cet espace qui expriment l'éventail de la polysémie d'une unité lexicale (qui sont le reflet de la richesse

sémantique d'une unité lexicale). Nous décrivons dans la suite la construction de ces « régions » de l'espace sémantique.

Pour Ploux et Victorri (1998b, p. 3), « *deux unités lexicales sont en relation de synonymie si toute occurrence de l'une peut être remplacée par une occurrence de l'autre dans un certain nombre d'environnements sans modifier notablement le sens de l'énoncé dans lequel elle se trouve* » (1998b, p. 3). Ils dégagent deux classes de synonymes, selon qu'ils sont « purs » ou « partiels ». Les synonymes purs sont les synonymes en relation « réflexive, symétrique et transitive » (relation d'équivalence). Les auteurs considèrent qu'il existe très peu de telles relations. De fait, on assimilera la synonymie à une synonymie partielle.

La synonymie partielle est une relation réflexive, symétrique et partiellement transitive. La réflexivité est une propriété assez évidente à vérifier : n'importe quelle unité lexicale peut se remplacer par elle-même dans n'importe quel contexte. Pour la symétrie de la relation, Manguin (2004), en se basant sur les travaux de Kahlmann (1975), la définit empiriquement comme une relation du type : *si A peut remplacer B (sans changer la portée du message) dans au moins un contexte donné, alors B peut remplacer A (sans changer la portée du message) dans au moins un autre contexte.* Enfin, la transitivité partielle ouvre la voie à de nombreuses unités lexicales candidates et pose un problème de limite pour relier deux unités lexicales entre elles par une relation de synonymie : jusqu'où une unité lexicale est-elle synonyme d'une autre ?

Par exemple, nous avons la chaîne suivante :

$$fenêtre \rightarrow baie \rightarrow golfe$$

or

$$fenêtre \nLeftrightarrow golfe$$

L'auteur propose donc d'utiliser « l'indice de similitude », défini pour deux unités lexicales A et B par :

$$S_A = \frac{card(syn_A) \cap card(syn_B)}{card(syn_A)} \qquad \text{(équation 2),}$$

et

$$S_B = \frac{card(syn_A) \cap card(syn_B)}{card(syn_B)} \qquad \text{(équation 3).}$$

Deux unités lexicales associées par une relation de transitivité sont considérées synonymes si les valeurs de leur indice de similitude relatif sont supérieures à 0,5. Afin d'éviter la situation aberrante de l'exemple précédent, qui correspond à deux unités lexicales d'indice de similitude 0,5 parce qu'elles n'ont qu'un seul synonyme en commun et qu'un seul synonyme (en plus d'elles-mêmes), l'auteur propose d'ajouter la condition suivante : un mot peut être relié à un autre s'il possède au moins deux synonymes (en plus de lui-même), et que ceux-ci sont communs aux deux unités à relier.

1.1.2.1 Le graphe de synonymes

Le Dictionnaire Electronique des Synonymes (DES) est construit à partir de la théorie des graphes de synonymie, dont la construction est décrite par Morel et François (2015). Les sommets constituent les entrées (mots) et les arêtes les liens synonymiques, identifiés à partir de la relation de synonymie partielle définie précédemment. Le corpus de connaissances du DES est constitué des listes de synonymes extraites de sept dictionnaires français : le Bailly, le Benac, le Du Chazeaud, le Guizot, le Lafaye, le Larousse, et le Robert. Cette base de données a été « symétrisée » (Morel and François, 2015), ce qui permet d'augmenter grandement le nombre de liens entre unités lexicales considérées comme synonymes quand bien même les dictionnaires sources ne les auraient pas considérés comme tels.

L'ensemble des sources, s'il est conséquent, est néanmoins fini et l'analyse de l'espace sémantique ne peut donc prétendre à une exhaustivité. Afin de pallier en partie cette limite, des suggestions d'utilisateurs sont intégrées et, depuis 2013, un système probabiliste repère et propose des liens synonymiques manquants. En conséquence, le corpus évolue désormais en dehors des stricts référentiels académiques vers un corpus d'usages dans la langue. En 2015, le DES contenait 49 254 entrées et 202 776 relations synonymiques réciproques (Morel and François, 2015).

La figure 4 est obtenue avec l'entrée « accident » dans le DES. Les arêtes représentent la relation de synonymie entre chaque sommet. Il s'agit toutefois ici d'une illustration : tous les synonymes et toutes les relations synonymiques de l'unité lexicale ne sont pas représentés.

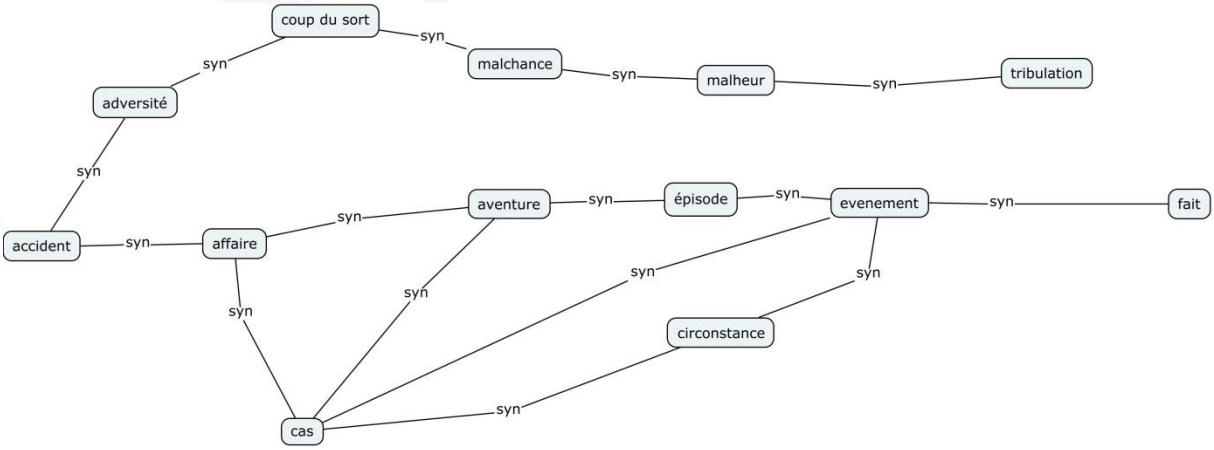


Figure 4 : un exemple de graphe des synonymes

Une fois le graphe élaboré, la recherche de groupes de synonymes d'une même unité lexicale est effectuée. Pour représenter ces groupes de synonymes d'une même unité lexicale, on utilise la notion de clique, soit « *un ensemble le plus grand possible de sommets du graphe tous reliés deux à deux* » (Ploux and Victorri, 1998a, p. 7). Cela permet de présenter le caractère monosémique (i.e. tous les synonymes de l'unité lexicale ont le même sens), homonymique (l'ensemble des synonymes est séparable en au moins deux sous-ensembles disjoints pour la relation de synonymie), ou polysémique (ni monosémique ni homonymique) de l'ensemble des unités lexicales associées par une relation de synonymie. A partir d'une unité lexicale, on parcourt le réseau formé de proche en proche jusqu'à rencontrer la limite correspondant à une

unité lexicale qui n'est pas synonyme avec chacune des unités lexicales précédentes. Le processus est itératif et se poursuit en partant de la première unité lexicale exclue de la clique précédente. Une clique est un sous-graphe complet maximal.

La figure 5 représente le graphe de synonymie et les quatre premières cliques proposées par le DES pour l'entrée « accident », soit :

- accident, adversité, coup du sort, malchance, malheur, tribulation ;
- accident, affaire, aventure, cas, évènement, fait ;
- accident, affaire, aventure, épisode, évènement, fait ;
- accident, affaire, cas, circonstance, évènement, fait.

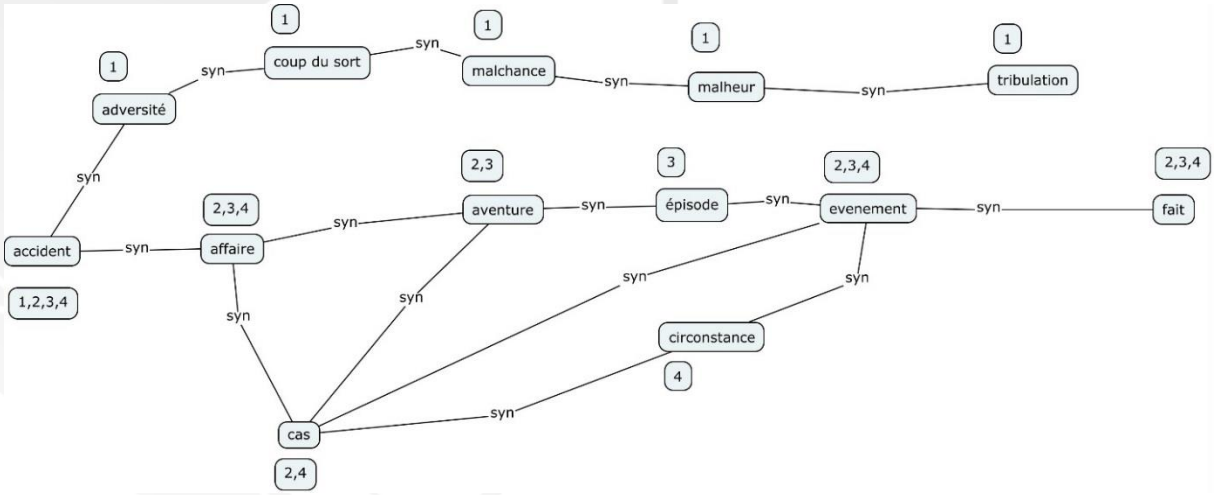


Figure 5 : un exemple de cliques de synonymes

1.1.2.2 Construction de l'espace sémantique de « accident »

Pour construire un espace sémantique, nous nous plaçons dans l'espace euclidien engendré par les synonymes et nous faisons correspondre à chaque clique un sommet d'un hypercube de cet espace *: « si l'on appelle u_1, u_2,..., u_n les synonymes, et c_1, c_2,..., c_p les cliques associées à l'unité étudiée, le synonyme u_i correspond au $i^{ème}$ vecteur de base de cet espace, et la clique c_k à un point dont les coordonnées x_{ki} valent 0 ou 1 suivant que le synonyme correspondant appartient ou non à la clique* » (Ploux and Victorri, 1998b, p. 12).

De là, les auteurs proposent de définir une métrique entre cliques, le χ^2, telle que :

$$\chi^2 = d^2(c_k, c_l) = \sum_{i=1}^{n} \frac{x}{x_i}\left(\frac{xk_i}{x_k} - \frac{x_{li}}{x_l}\right)^2 \qquad \text{(équation 4)},$$

avec $x_i = \sum_{j=1}^{p} x_{ji}$, $x_k = \sum_{i=1}^{n} x_{ki}$ et $x = \sum_{j=1}^{p}\sum_{i=1}^{n} x_{ji}$.

On remarque que chaque synonyme intervient dans le calcul avec un « poids » plus faible si le synonyme est présent dans un grand nombre de cliques et le point représentant une clique est d'autant plus proche de l'origine que la clique comporte de synonymes. Le problème sous-jacent de cette approche est le très grand nombre de dimensions de l'espace généré (égale au nombre de synonymes) et qui ne permet pas une exploitation aisée par l'humain.

Afin de réduire la dimension du problème, nous procédons à une Analyse par Composantes Principales (ACP), en nous appuyant sur le cours de Duby et Robin

(2006). L'ACP est une méthode descriptive multidimensionnelle de type factorielle. L'objectif est de déterminer des dimensions nouvelles et non corrélées qui décrivent le mieux notre objet d'étude, en l'espèce un hypercube, particulièrement lorsque celui-ci est au départ dimensionné dans un espace à n dimensions avec n élevé. A partir d'une matrice comportant p valeurs exprimées en n dimensions, on construit une « approximation » géométrique par projection sur un espace de dimension $q << n$. Dans notre cas, on cherche à décrire les cliques de synonymes selon de nouvelles dimensions issues des vecteurs initiaux que sont les synonymes. Pour cela, il nous faut trouver un système d'axes et de plans tels que les projections des nuages de points (les observations) sur ces axes et ces plans « *permettent de reconstituer les positions des points les uns par rapport aux autres, c'est-à-dire avoir des images les moins déformées possible* » (2006, p. 5).

Toutefois, contrairement à ce qui se fait dans le cadre d'une ACP classique, la distance entre deux points retenue pour analyser l'espace sémantique n'est pas la distance euclidienne, mais la métrique χ^2 (Ploux and Victorri, 1998b).

En outre, il est opportun de choisir une origine autre que celle correspondante au vecteur de coordonnées nulles pour centrer le nuage de points ; la nouvelle origine correspond donc au centre de gravité du nuage de cliques, calculé en attribuant la même importance à toutes les variables au départ. L'ensemble des cliques est exprimé dans ce nouveau système de coordonnées.

La matrice correspondante est ensuite multipliée par sa transposée pour obtenir une matrice symétrique donc diagonalisable. Cela permet de rechercher des axes d'inertie, i.e. des axes tels que la projection du nuage sur ces axes ait une variance maximale. On change donc de repère pour diagonaliser la matrice : par définition, chaque valeur sur la diagonale de la matrice M obtenue est la variance de la population selon chaque axe de notre nouveau repère.

Nous venons d'expliciter la méthode de l'ACP utilisée par le dictionnaire des synonymes pour réduire l'espace sémantique. L'espace sémantique est ainsi ramené à un ensemble de dimensions – dont la portée sémantique est ordonnée en fonction des valeurs propres de la matrice M – qui permettent d'expliquer la richesse polysémique d'une unité lexicale à travers les relations entre ses synonymes et leurs utilisations dans des corpus de documents. Visuellement, les régions de l'espace sémantique sont représentées en projection sur des plans dont les axes sont orientés pour marquer les oppositions de sens. Il s'agit par la suite d'identifier une représentation convenable ainsi que les axes correspondants.

1.1.3 Représentations de l'espace sémantique de l'unité lexicale « accident »

L'unité lexicale « accident » comporte soixante-et-onze synonymes (en juin 2017)[9]. Il y a cent-et-une cliques[10]**.** Le classement des synonymes proposé par le DES est par ordre décroissant, en fonction du rapport entre le nombre de cliques auxquelles il appartient et le nombre total de cliques. L'unité lexicale « évènement » est celle qui partage le plus de cliques avec l'unité lexicale « accident » ; elle est considérée comme le synonyme le plus représentatif de l'unité lexicale dans le corpus considéré.

1.1.3.1 Les dimensions de « l'accident »

Le plan formé par les deux dimensions les plus explicatives de la sémantique de l'accident est représenté à la figure 6 :

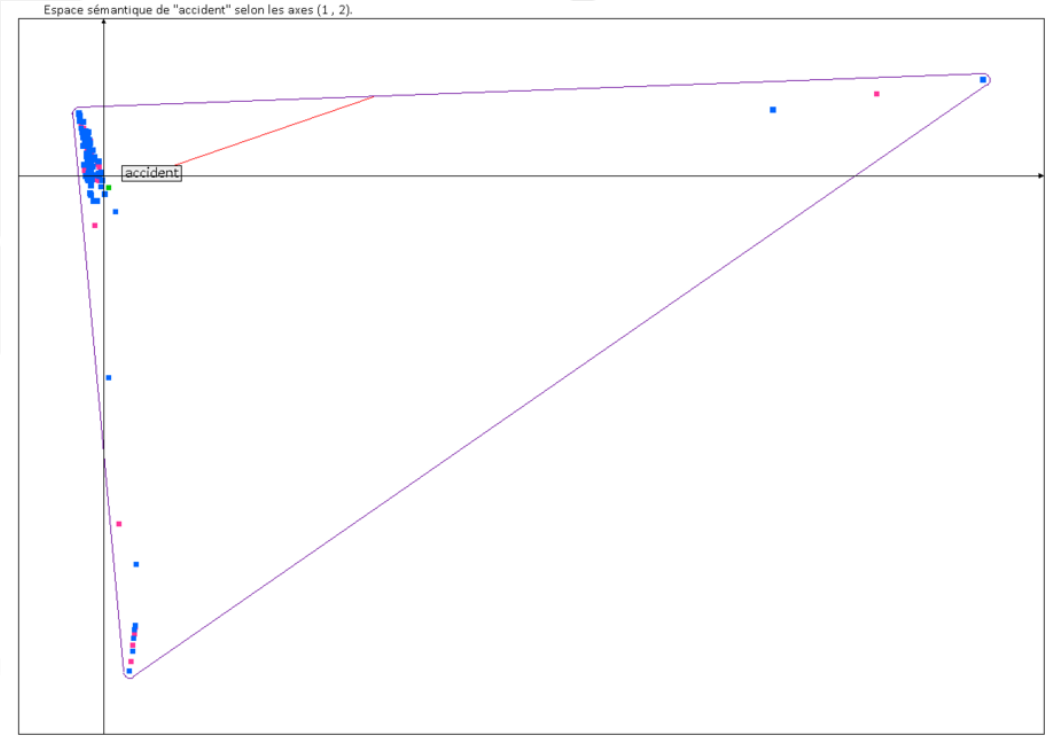


Figure 6 : projection de l'espace sémantique de l'unité « accident » selon les deux premiers axes d'analyse sémantique

Les points bleus sont les projections de cliques ; un point rouge correspond à la projection du centre de gravité de l'espace sémantique d'un synonyme. Le point vert, ici « accident », est le centre de gravité de l'espace sémantique de l'unité lexicale regardée, lui-même délimité par le contour violet. Victorri (2002), dans un document

9 Par ordre alphabétique : à-coup, accroc, accrochage, adversité, affaire, aléa, anicroche, artéfact, aspérité, attribut, avatar, aventure, bûche, calamité, capotage, cas, cataclysme, catastrophe, chagrin, choc, chute, circonstance, collision, complication, conséquence, contretemps, coup, coup du sort, coup dur, dénivellation, dérapage, drame, effet, embardée, ennui, épisode, évènement, explosion, fait, hasard, incidence, incident, inégalité, irrégularité, lésion, malchance, malencontre, malheur, manque de pot, méchef, mésaventure, montagne, mouvement, mouvement de terrain, occasion, panne, pépin, péripétie, phénomène, pli, plissement, relief, repli, résultat, revers, tête-à-queue, tonneau, tragédie, tribulation, tuile, vicissitude. Le classement des premiers synonymes est le suivant : évènement, incident, mésaventure, tribulation, aventure, malheur, circonstance, anicroche, hasard, avatar, malchance, pépin, cas, épisode, fait et vicissitude.
10 Présentées en annexe.

préliminaire, propose des pistes d'interprétation de ce type de graphe. La méthode repose en bonne partie sur l'intuition. Nous allons essayer de livrer notre interprétation en suivant cette méthode, puis nous pointerons les limites d'une telle représentation.

Trois groupes se détachent : le groupe principal, très dense, proche du centre du repère et grossièrement orienté selon l'axe vertical ; un deuxième, moins dense, en bas, plutôt orienté verticalement également ; enfin, un troisième, formé de trois points alignés et espacés, en haut à droite orienté à peu près horizontalement.

Par construction de l'ACP, les projections éloignées de l'origine sont moins fiables dans leur représentativité du nuage de points et il convient donc de les écarter : en l'espèce, nous ne retenons pas les deux projections de cliques en haut à droite, issues de l'unité lexicale « inégalité » (dans le sens de la rugosité d'une surface, d'un niveau).

Par ailleurs, les projections des cliques et des synonymes dans le bas de la figure 6 sont probablement peu significatives quant à la construction de l'Axe 2, compte tenu de leur positionnement par rapport à l'origine et leur nombre relativement réduit.

Nous resserrons notre interprétation autour du groupe principal et nous regardons les sept premiers synonymes proposés par le dictionnaire, cf. figure 7 :

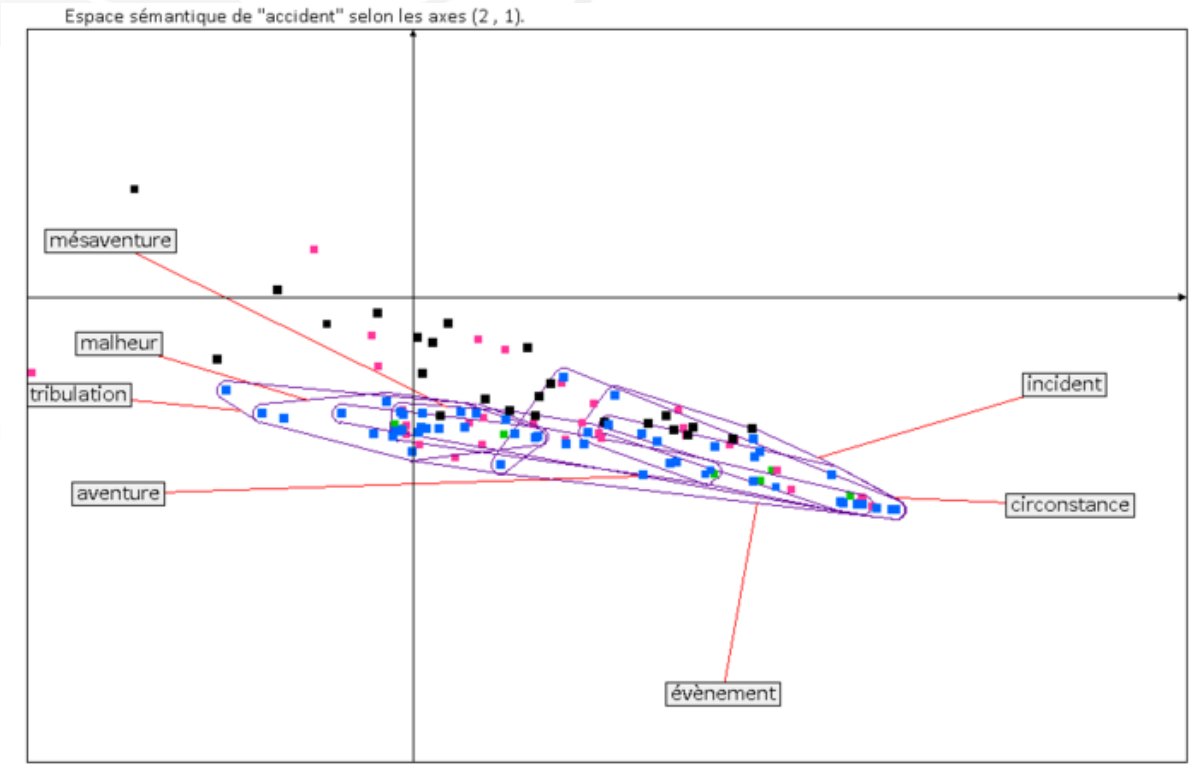


Figure 7 : plan 2,1 de l'« accident »

Par rapport à la figure 6, la figure 7 est exprimée dans le repère (0, Axe 2, Axe 1) après rotation des axes. Les espaces sémantiques des synonymes semblent très similaires, en forme et en orientation. Ils recouvrent la grande majorité des cliques. Les espaces sémantiques de « évènement », « incident » et « circonstance » sont à droite de l'axe vertical ; on en déduit qu'ils pourraient vraisemblablement être utilisés de manière interchangeable sans ambiguïté comme synonymes de « accident » (par rapport à la dimension de référence, en l'espèce l'Axe 2 qui reste à déterminer). Au contraire, les

espaces sémantiques de « mésaventure », « malheur » et « tribulation » sont à cheval sur cet axe : leur usage synonymique (selon la signification portée par l'Axe 2) introduit une ambiguïté (un sens potentiellement contraire).

Nous regardons maintenant les deux cliques les plus opposées soit :

- à gauche de l'axe : accident, revers, tribulation, vicissitude ;
- à droite de l'axe : accident, affaire, circonstance, fait, épisode, évènement.

En se basant sur les définitions données par le CNRTL des unités lexicales composant ces cliques, se dégage un critère de différenciation de sens : à droite de l'axe, le sens marque une concrétisation de l'accident dans le monde physique tandis qu'à gauche, l'accident est envisagé de manière possible, imaginaire ; l'opposition est de l'ordre de l'antagonisme objectif / subjectif.

Si l'on centre l'analyse sur l'espace sémantique d'« évènement » (premier synonyme), en regardant les deux cliques les plus éloignées on obtient la figure 8 :

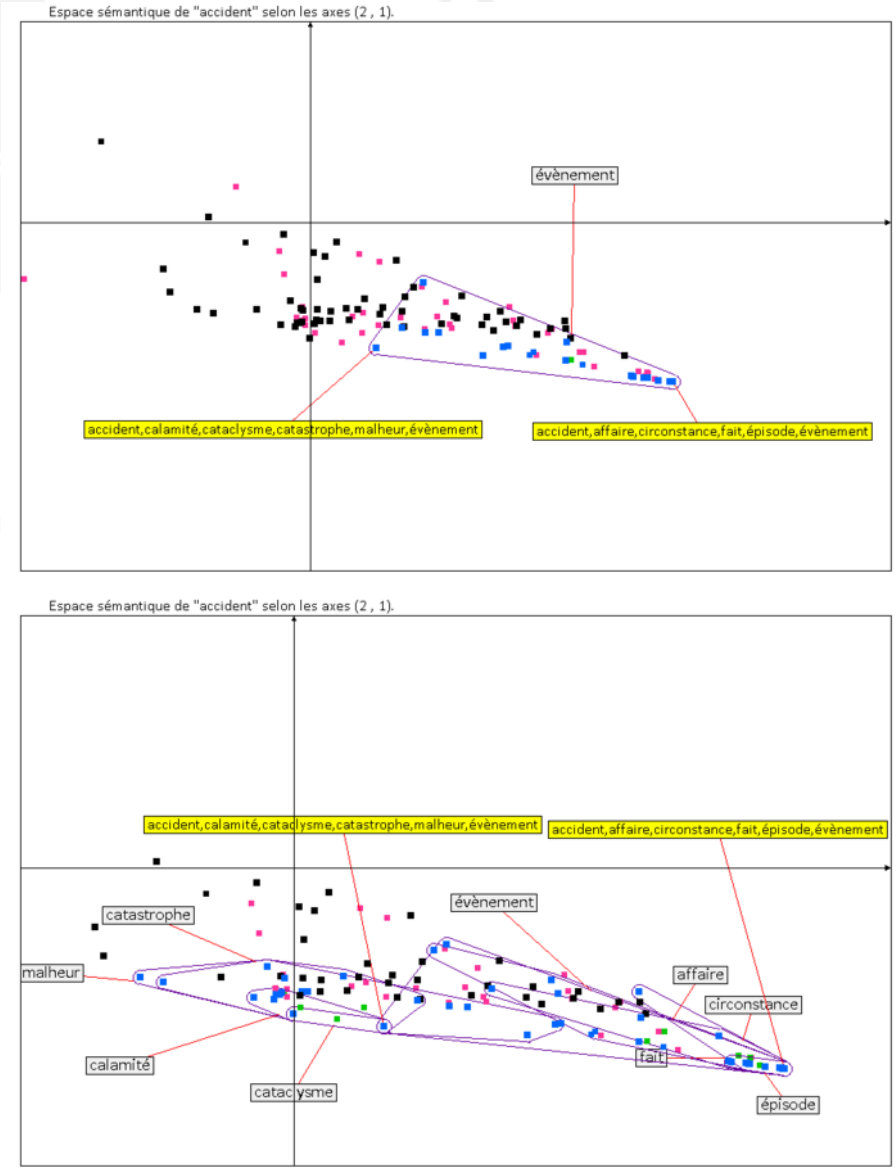


Figure 8 : les espaces sémantiques des cliques les plus éloignées d'« évènement »

Au vu de la forme et de l'orientation de l'espace sémantique d'« évènement », nous pouvons penser qu'il contribue significativement à la construction de l'Axe 2 et donc que son influence sur la sémantique de l'accident est forte. L'usage d'« évènement » comme synonyme d'accident renvoie à la manifestation de l'accident dans le monde physique au travers de ses conséquences. Cela vient renforcer l'acception commune de la définition de l'accident (définition II) tout en apportant une précision : le sens varie selon l'Axe 2 en fonction de « l'espace » qu'occupe l'évènement. S'il est circonscrit, l'accident s'apparente à une « affaire » ou « circonstance », un « épisode » ; s'il sature l'existant, il devient synonyme de « catastrophe », « d'adversité », de « cataclysme ». L'évolution en fonction de ce critère est précisée par la projection sur les Axes 3 et 4, dans la figure 9 ci-dessous :

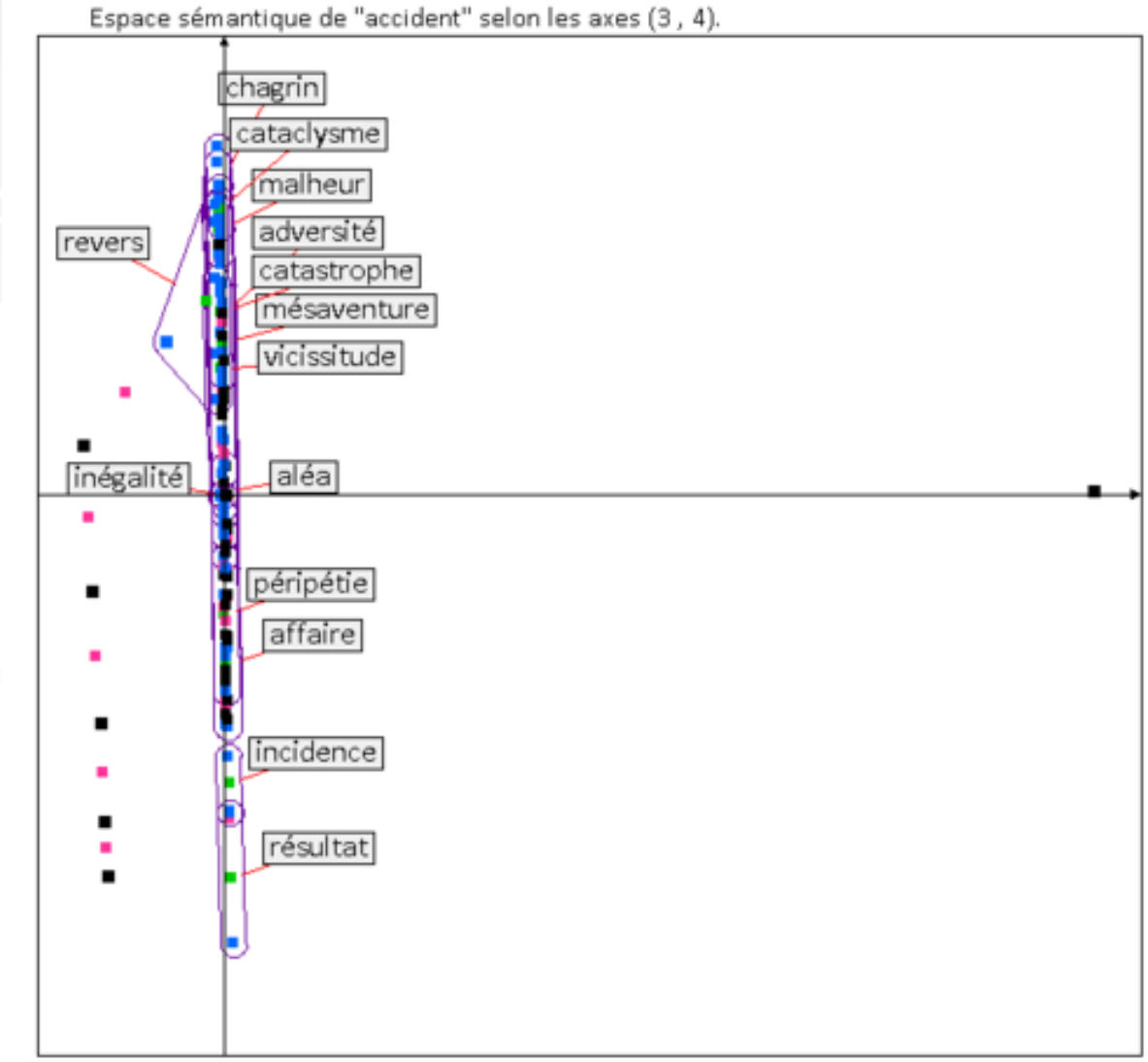


Figure 9 : un autre plan, les dimensions 3,4

On observe une distribution très étagée des cliques et des synonymes selon une droite presque parfaitement verticale selon la dimension 4. On élimine la dimension 3 de notre analyse car elle n'est constituée que de peu de projections très éloignées de l'origine et qui ne sont donc probablement pas significatives.

Nous pouvons schématiser cette variation de sens à l'aide du schéma conceptuel de noyau de sens de « accident » (cf. section 1.1 *supra*) : dans le cas de l'accident

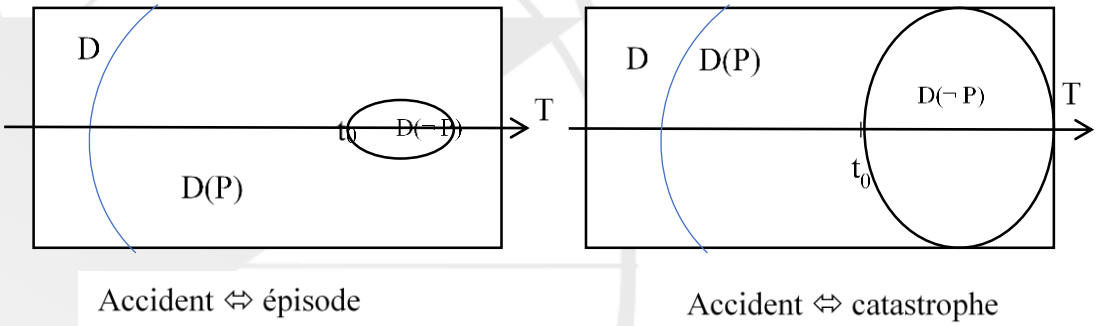


Figure 10 : variation du sens d'« accident » en fonction des frontières rencontrées de D(P)

« épisode » (ou circonscrit dans ses conséquences), la trajectoire T peut contourner les limites du domaine D(P), tandis que l'accident « catastrophe » sature le domaine d'existence (Figure 10) :

En revenant à la figure 9, on trouve en outre au plus proche de l'origine, l'unité lexicale « aléa », qui serait donc la plus neutre selon la dimension *circonscrit/saturant* que nous venons d'identifier. En appui de cette considération, notons qu'en France dans le cadre de l'élaboration des Plans de Prévention des Risques Technologiques (PPRT), l'unité lexicale « aléa » est utilisé comme ceci : « *L'aléa technologique est une composante du risque industriel [...]. Il désigne la probabilité qu'un phénomène dangereux produise, en un point donné du territoire, des effets d'une intensité physique définie.* » Avec la notion de probabilité, quantitative, et des effets « définis », la définition proposée par le PPRT est neutre, et s'affranchit de l'ampleur du phénomène comme critère de définition.

1.1.3.2 La nuance sémantique entre deux unités lexicales par leurs cliques respectives

Intéressons-nous maintenant aux subtilités des synonymies partielles entre les unités lexicales. Quelles sont les différences de sens entre les deux premiers synonymes « d'accident » respectivement « évènement » et « incident » ? Il y a vingt-six cliques dans l'espace d'« évènement » et dix-huit cliques dans celui d'« incident ». Ces deux espaces sémantiques ont dix cliques en commun ; les indices de similarité relatifs sont :

$$S_{évènement} = \frac{10}{26} = 0{,}38$$

et

$$S_{incident} = \frac{10}{18} = 0{,}56$$

« Incident » est plus influencé par « évènement » que le contraire. Regardons maintenant $E_{évènement} \cap E_{incident}$. Sans tenir compte de leur apparition dans les différentes cliques, nous avons compté et classé par fréquence d'apparition les synonymes.

On trouve les résultats suivants :

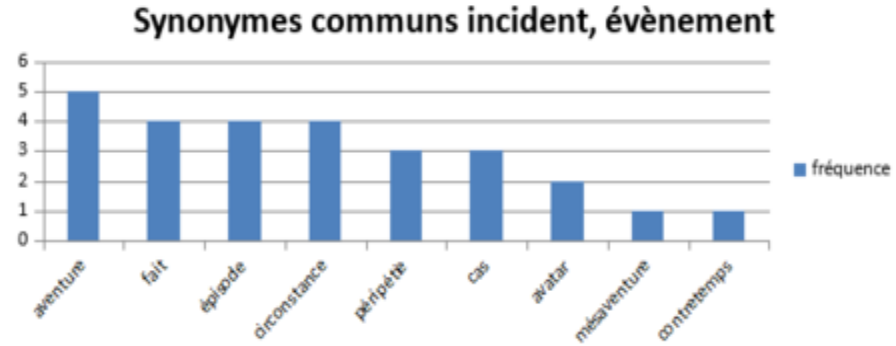


Figure 11 : l'intersection entre cliques « évènement » et « incident »

Regardons maintenant les cliques qui appartiennent à l'espace « évènement » sans appartenir à l'espace « incident » (Figure 12) :

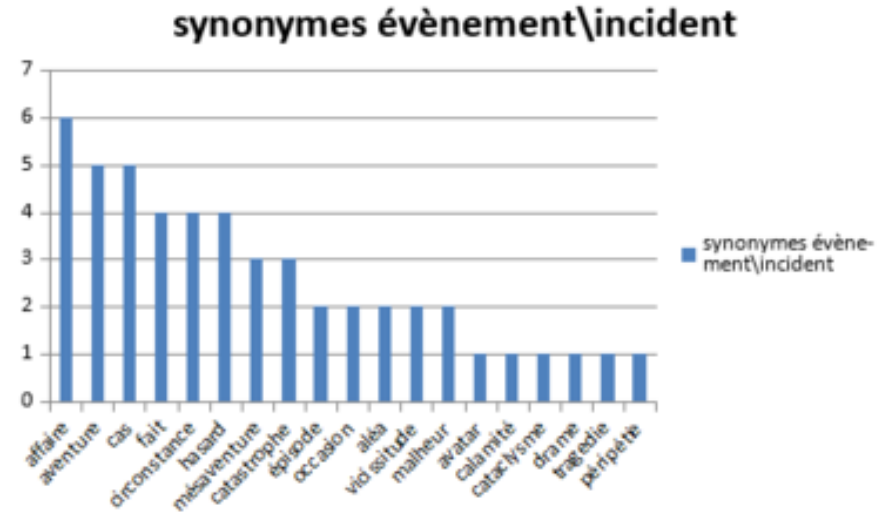


Figure 12 : cliques de synonymes d'« évènement » sans intersection avec « incident »

Puis, pour les huit cliques « incident » sans « évènement » (Figure 13) :

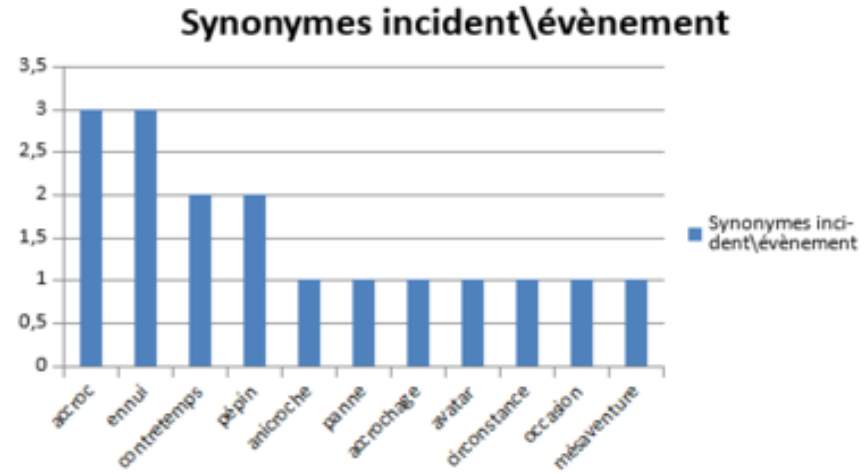


Figure 13 : cliques « incident » sans « évènement »

Pour affiner la différenciation sémantique entre « évènement » et « incident », on propose un tri selon les règles suivantes : on élimine les synonymes les moins discriminants, c'est-à-dire ceux qui sont présents à la fois dans l'intersection strictement et dans au moins un espace sémantique moins l'intersection. De là, à partir des fréquences calculées auparavant, on peut déduire, par fréquence d'apparition dans les cliques respectives que :

- « évènement » est sémantiquement influencé par « affaire », « hasard », « aléa », « malheur », « occasion », « vicissitude », « calamité », « drame », « tragédie » ;
- « incident » est sémantiquement influencé par « accroc », « ennui », « pépin », « accrochage », « anicroche », « occasion », « panne ».

« Evènement » semble être porté par la notion d'absence de contrôle, de « coup du sort ». Appréhendé socialement ; on peut même y percevoir une dimension épique dans la « traîne » sémantique. Il y a aussi une part de fatalisme, de quelque chose qui dépasse l'entendement humain. Présente, la connotation négative est moins marquée que le caractère imprévisible de la survenance. D'ailleurs « vicissitude » ou « occasion » s'emploient aussi pour parler d'évènements pouvant être heureux. L'unité lexicale « affaire » est symptomatique de l'idée d'intérêt, que le collectif « s'empare » de ce qu'il vit, mais qu'il peut aussi être dépassé, submergé.

« Incident » marque une appropriation plus individuelle et circonscrite. Le préfixe *ac* apporte la notion de rapprochement, de contact et le préfixe *an* la notion de tourner autour (« accrochage » et « anicroche » sont aussi utilisés dans le lexique militaire où la violence est bien présente et le caractère « non attendu » aussi). On voit même faiblement l'apparition d'un cas « d'incident » très instancié dans le monde physique, la « panne ».

Nous pourrions continuer à fournir une interprétation de plus en plus détaillée de la sémantique de l'accident, mais il faudrait pouvoir se garantir des biais inhérents à cette représentation que le CRISCO nous fournit, notamment concernant l'ACP : Le nombre d'axes à analyser est la première des limites. En effet, il s'agit de trouver un compromis entre une maximisation de l'explication de la variance et une réduction du nombre de dimensions souhaité. Pour cela il faudrait avoir accès au code du logiciel. S'il existe des méthodes pour aider à choisir le nombre de dimensions adapté, dans notre cas il n'y a aucun indicateur et il vaut mieux ne pas aller au-delà des premières dimensions d'explications sous peine d'errements. En particulier, lorsqu'on effectue une ACP, il est indispensable de regarder le carré du cosinus de l'angle entre le vecteur d'un individu et l'axe de projection en question, afin d'estimer la qualité de la projection. Comme nous n'avons aucune indication à ce sujet, il prévaut de s'intéresser en priorité aux individus proches de l'origine du plan en question et de se garder d'interpréter trop rapidement les individus qui en sont éloignés. En ayant accès à ces données supplémentaires, il serait intéressant de se pencher sur des dimensions moindres, mais qui cependant pourraient venir compléter la sémantique de l'accident sur sa polysémie.

*

* *

Dans cette section, nous avons foré dans les corpus linguistiques selon différentes approches pour mieux cerner la structure de l'espace sémantique d'« accident » autour de son noyau de sens et préciser son acception commune. A ce stade, dans une perspective de gestion des risques industriels, nous définissons l'« accident » comme suit :

> ***L'accident caractérise une rencontre avec une frontière d'un domaine de « normalité » ; il se rapporte à une souffrance, et n'existe que par les effets destructeurs, notamment physiques, subis ou ressentis par un individu ou un groupe dans le nouveau domaine d'existence où ce dernier (individu ou groupe) est subitement projeté.***

L'évaluation des conséquences de l'accident et de l'enveloppe du domaine de normalité sera bien entendu différente selon que l'observateur subit lui-même l'accident ou est « extérieur ». L'accident est un concept anthropocentrique, un « fait » social. L'instanciation de l'évènement « accident » peut être destinée à lui apporter une « explication » rationnelle. Dans tous les cas, l'accident s'inscrit dans un processus dynamique, un changement d'état pénible. Il est détaché de la notion de durée en lui-même, contrairement à la trajectoire à laquelle il participe, qu'il s'agisse des actes qui y ont conduit ou de la réaction humaine pour mettre fin à ses conséquences.

Cette section nous a permis d'introduire des concepts d'analyse sémantique qui seront réutilisés au chapitre 3. Dans la suite de ce chapitre, nous proposons de mieux comprendre comment l'accident est pris comme objet d'étude, à travers le cas de *Deepwater Horizon*, dont les conséquences ont conduit l'Etat américain à mettre en œuvre des moyens d'ingénierie considérables.

La définition de l'accident renvoie à une transition : ce dernier n'a pas d'existence propre autrement qu'attaché aux conséquences qu'on lui prête et qui contrastent avec une « normalité », difficile à appréhender. La crise qui s'ensuit traduit un moment de reconfiguration dans un nouveau domaine d'existence, peu exploré en tant que tel. Nous allons montrer que pour l'essentiel, les études « objectivent » l'accident en montrant que ce qui est vécu comme « accident-catastrophe » n'est qu'un « accident-épisode » que la trajectoire aurait pu éviter (cf. Figure 10 *supra*). Le naufrage de la plateforme *Deepwater Horizon* et la gestion de ses conséquences n'échappent pas à cette règle.

1.2 Des données de l'accident de *Deepwater Horizon*

Dans cette section, nous montrons l'étendue des données disponibles sur le cas *Deepwater Horizon* et la problématique de collecte que cette étendue soulève. Nous conservons une approche « fouille de données », basée sur l'exploration systématique de l'Internet. Rappelons tout d'abord l'accident dans ses grandes lignes.

Le puits était le premier prévu sur la concession *Macondo*, nom de code donné par BP à la concession Mississippi Canyon Block 252 dans le Golfe du Mexique,

considérée alors comme un « *golden block* », un bloc à très fort potentiel de production. Le 20 avril 2010, le puits entre en éruption après la perte de contrôle de l'équipe de forage à la fin des tests de « complétion »[11]. Un « *blowout* », une venue de gaz et de boue sur le pont de la plateforme, survient aux alentours de 21 h 30 et provoque deux explosions : la première à 21 h 49, la seconde dix secondes après. En effet, le Bloc Obturateur de Puits, ou BOP (*BlowOut Preventer*)[12] ne s'est pas fermé à temps pour isoler mécaniquement la plateforme du puits. La montée du fluide, composé majoritairement de gaz inflammables et de boue, a atteint le « *riser* »[13], jusqu'au niveau du pont de la plateforme. L'évacuation de la plateforme a été décidée, et l'équipage a été principalement secouru par un navire de service qui était à proximité du *rig* alors qu'il effectuait un transfert de boue (un matériau de forage) au moment du *blowout*. A 22 h 04, une opération de recherche et de sauvetage des survivants est lancée par les garde-côtes américains. La plateforme, en flammes, coule le matin du 22 avril 2010. Onze travailleurs sont décédés.

Le 21 avril, BP et Transocean envoient des Robots Sous-Marins (*Remote Operated Vehicles*, *ROV*), sur le fond marin, pour tenter de fermer le BOP, par plus de 1 500 m de profondeur, afin de stopper l'alimentation en combustible de l'incendie. Toutes les tentatives pour enclencher la fermeture du BOP, y compris à bord de la plateforme, demeurent sans succès. L'organisation des secours, au départ destinée à porter assistance aux naufragés et à lutter contre l'incendie, va changer profondément de nature et d'ampleur à la découverte des fuites sous-marines.

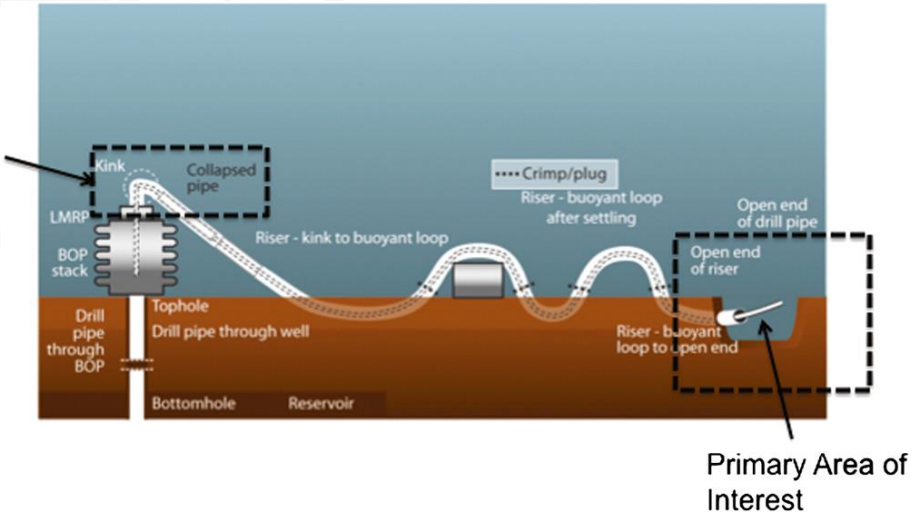


Figure 14 : schéma des fuites sous-marines, d'après (McNutt et al., 2012c)

11 Il s'agit de sceller le puits avec du ciment pour le mettre en sécurité, puis de le tester en pression. Cette opération permet à la plateforme de forage de se déconnecter et de laisser la place par la suite à une plateforme de production, qui une fois connectée, « brisera le scellé » et commencera l'extraction.

12 Equipement technique pour les opérations de *well control*, utilisé à la fois en opérations de routine et en barrière de sécurité ; il est installé sur la tête de puits (Deepwater Horizon Study Group, 2011, pp. 25–28). Une description précise de cet organe critique est donnée dans le rapport de BP (Appendix H: description of the BOP Stack and Control System Bly and BP, 2010) du DNV (Det Norske Veritas, 2011, Pp. 14–17).

13 Le *riser* est le flexible qui relie la plateforme de forage au puits et dans lequel est descendue la tige de forage. Les fluides de contrôle du puits y circulent également pendant les opérations de forage.

En effet, dès le 23 avril, les ROV découvrent à l'extrémité du *riser*, arraché de la plateforme pendant sa chute, une première fuite ; puis une deuxième le lendemain, à la connexion du *riser* sur le BOP.

Le 29 avril 2010, le ministre de l'Intérieur Ken Salazar décrète la pollution de niveau *Spill Of National Significance*, *SONS*, le plus haut niveau en matière d'engagement national[14]. Jusqu'au 15 juillet 2010, la plus grande organisation montée en temps de paix par les États-Unis va fournir un effort sans précédent pour tenter de reprendre le contrôle d'un puits d'hydrocarbures sous-marin devenu « sauvage ».

1.2.1 Stratégie de collecte

Une telle crise a bien entendu donné lieu à d'innombrables documents, des plus techniques au plus politiques. Comment les explorer ? Dans un premier temps, nous avons pris connaissance du cas dans une démarche experte. Celle-ci s'est appuyée sur une collecte d'informations à partir de l'Internet. Nous pratiquons une Recherche Ouverte d'Information (ROI), dont « *la finalité principale (…) n'est pas la récupération ou l'accès à un document ou à un ensemble de documents pertinents par rapport à une interrogation déjà formulée, mais l'assistance à un utilisateur engagé dans une démarche d'enquête (…) à travers laquelle il sera amené à préciser simultanément les termes de sa demande et les ressources documentaires ou non-documentaires susceptibles de contribuer à y répondre* » (Zacklad, 2007, p. 2).

Bates (1989) propose un modèle de recherche d'informations en ligne et dans les systèmes d'information, qu'elle désigne par « *berry picking strategy* ». Ce modèle se veut analogue au comportement d'un humain lorsqu'il est à la recherche de données. En effet, l'auteur conteste le modèle « classique » de recherche d'informations centré sur l'adéquation entre la requête émise par le « chercheur » et la capacité d'une base de données à fournir la réponse appropriée, qui n'est effectivement plus adapté à la fouille de l'Internet. Le *berry picking* est structuré en fonction de :

- la nature de la recherche (requête) ;
- la nature du processus de recherche ;
- les techniques de recherche utilisées ;
- l'espace dans lequel la recherche est conduite.

Le modèle est centré autour des comportements du chercheur face à son travail. Ce travail est par nature évolutif. Les techniques de recherches proposées sont inspirées des travaux de Stoan (1984) et Ellis (1984, 1989). La plupart des requêtes sur le Web s'effectuent par mots-clés guidés par notre propre expertise et des concepts qui émergent pendant le processus de collecte de données. A partir de ces premiers éléments, la collecte est organisée en fonction des sources qui permettent de consolider et capitaliser l'information. Un classement *a priori* est appliqué tout au long du

14 Source : (*e-CFR: TITLE 40—Protection of Environment*, no date).

processus de collecte des données. Seules les données pertinentes selon notre univers d'intérêt sont collectées.

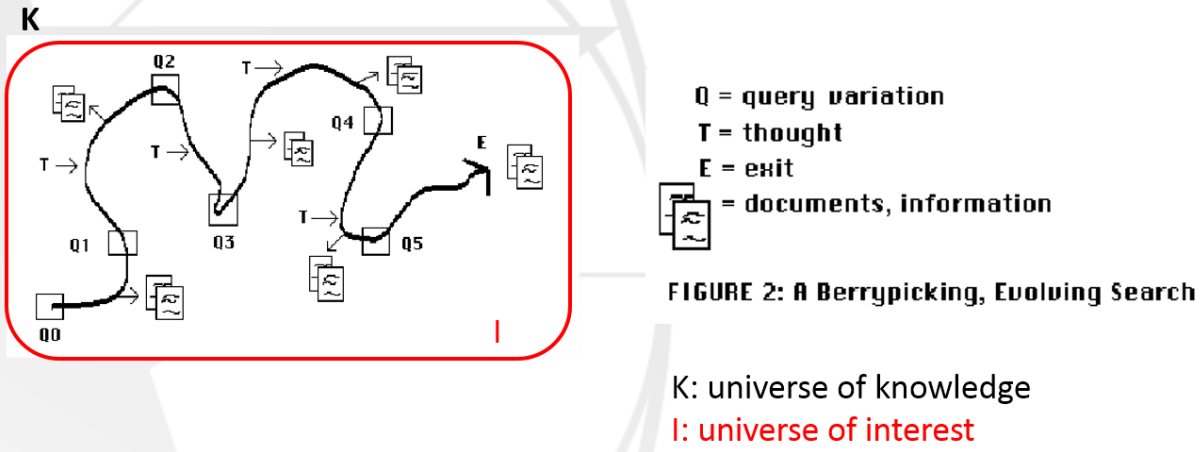


Figure 15 : schéma extrait de (Bates, 1989) illustrant le processus de recherche

Plusieurs stratégies permettent d'orienter la collecte :

- la chasse aux notes de bas de page permet de « remonter » vers les documents utilisés comme références dans l'ouvrage en question. Le chercheur s'affranchit ainsi de l'interprétation livrée dans l'ouvrage et élargit son corpus documentaire. De même, la chasse aux citations permet de « rebondir » vers des documents sources ;
- une fois que nous avons identifié une revue reconnue comme « référente » dans le domaine de recherche, notamment via son *impact factor* et le *board of editors*, nous allons cibler les années de publication les plus pertinentes. De la sorte, nous sommes certains d'avoir la totalité des informations disponibles dans cette revue sur le sujet en question. Cette technique exploite la loi de Bradford[15];
- la connexité ou « *area scanning* ». L'exemple donné par Bates est basé sur la collocation physique (dans une bibliothèque) d'ouvrages intéressants de prime abord (c'est-à-dire sélectionnés par le besoin de recherche) et d'un autre côté, la recherche d'ouvrages adjacents qui permettent d'élargir le champ de possibilités d'information autour du sujet de recherche. La connexité, dans la pratique d'aujourd'hui, est essentiellement portée par l'utilisation de mots-clés qui marquent comme de multiples étiquettes les documents. Mais il existe de très nombreux moyens quantitatifs pour mesurer la similarité de documents entre eux, donc de la possible connexité et *in fine*, de l'intérêt relatif. Le meilleur exemple pour illustrer ceci est la capacité qu'ont de très nombreux sites web et bases de données scientifiques à proposer des articles « en lien » avec l'article ou la page consultés ;
- la recherche par sujet dans les services de bibliographie (résumés et index). Aujourd'hui, les grands services de bibliographies ont indexé et classé la documentation par sujet. Il peut s'agir d'une manière de « rentrer » dans l'information ;
- par contraste avec la recherche par sujet, la recherche par auteur permet de retrouver l'entièreté de l'œuvre d'un auteur qui, *a priori*, a contribué à la recherche sur le sujet en question.

15 Source : http://statelem.com/

Ces techniques de recherche, en particulier celles par connexité et par indexation, ont bénéficié d'avancées considérables ces dernières années, notamment avec les moteurs de recherche, le lien hypertexte et le Web sémantique.

1.2.2 Les données accessibles par l'Internet

L'Internet est notre portail d'accès à la très grande majorité des données que nous récoltons à propos de l'accident de *Deepwater Horizon via* le moteur de recherche Google qui référence des sites web susceptibles de nous intéresser en fonction des requêtes qui lui sont soumises.

1.2.2.1 Abord de Deepwater Horizon via Google

Nous considérons intéressant *a priori* tout document qui mentionne l'évènement en utilisant les termes *Deepwater Horizon* ou *Macondo well* et en priorité aux données textuelles contenues dans des documents écrits en langage naturel, principalement en anglais. Notre stratégie de collecte, repose sur les hypothèses et choix suivants :

1. le cas auquel nous nous intéressons est *a priori* identifiable à l'aide des expressions « *Deepwater Horizon* », le nom de baptême de la plateforme de forage ou « Macondo », le nom de code de la concession pétrolière ;
2. nous traitons le cas *Deepwater Horizon* par le prisme de l'accident tel que nous l'avons exploré dans la première partie de ce chapitre ;
3. nous écartons les sources de données générées par la fiction sortie en 2016 ;
4. nous donnons notre premier intérêt aux sources officielles, institutionnelles ou gouvernementales ainsi qu'aux sources académiques et scientifiques[16].

Le tableau en page suivante recense le nombre de résultats donnés par Google en fonction des requêtes que nous lui avons soumises :

16 Nous expliquons *infra* notre approche pour l'évaluation de la qualité de l'information.

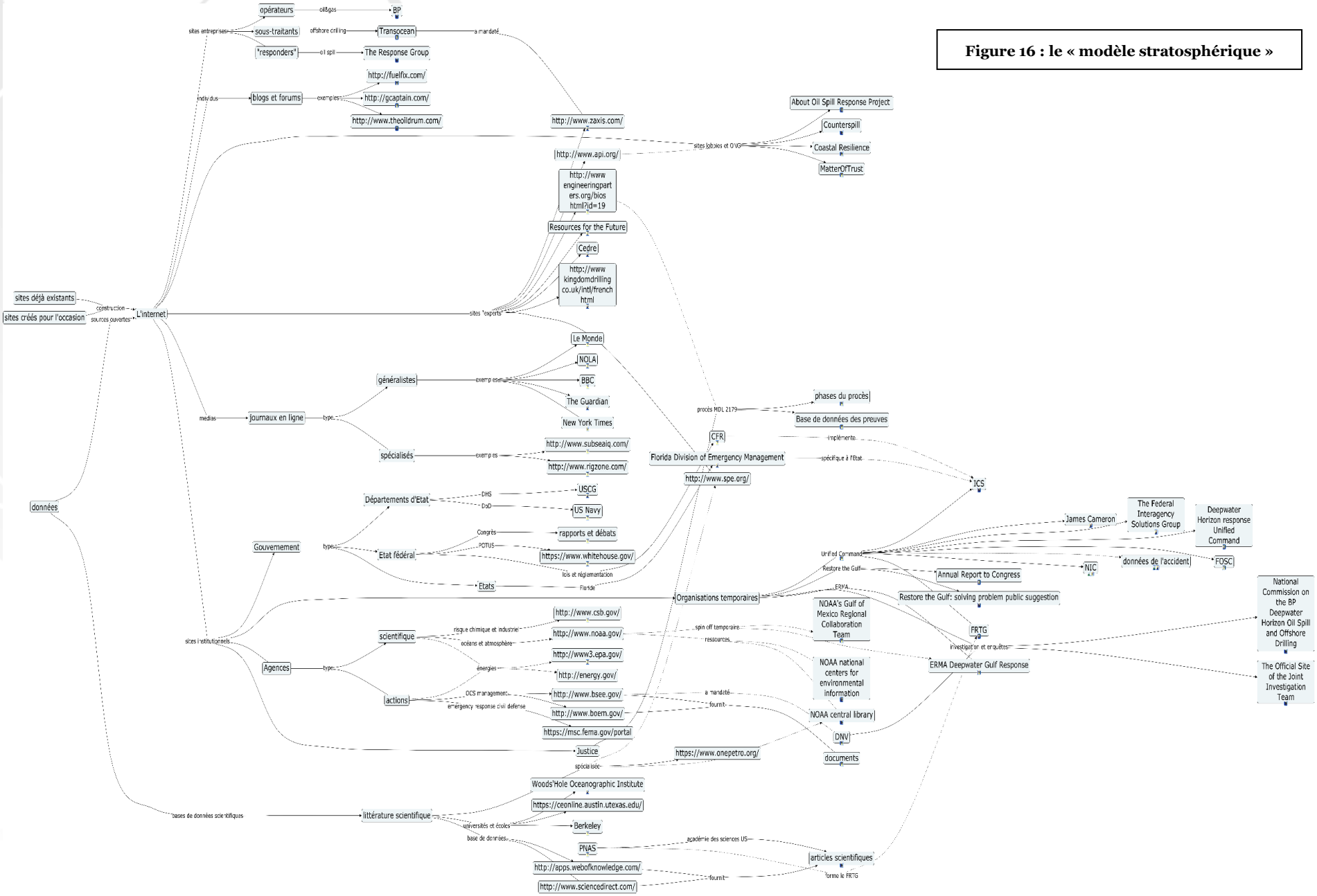
Google	Opérateurs	Type de fichier	Nombre de résultats (ordre de grandeur)
« deepwater horizon »	Terme exact	All	3 640 000
macondo « deepwater horizon »	ET, terme exact	pdf	117 000
		All	336 000
macondo OR « deepwater horizon »	OU, terme exact	All	7 470 000
macondo OR « deepwater horizon » – movie	OU, terme exact, SANS	All	2 780 000
macondo OR « deepwater horizon » site : https://www.sciencedirect.com/	OU, terme exact, site spécifique	All	Environ 4 460 résultats
macondo OR « deepwater horizon » site : https://wikipedia.org/	OU, terme exact, site spécifique	All	10 600
macondo OR « deepwater horizon » AND “accident”	OU, ET, terme exact	All	390 000
		Pdf	48 800
macondo OR « deepwater horizon » AND accident site : https://www.rigzone.com	OU, ET, terme exact, site spécifique	All	216
macondo OR « deepwater horizon » AND investigation	OU, ET, terme exact	All	371 000
macondo OR « deepwater horizon » site : https://www.energy.gov/	OU, terme exact, site spécifique	All	193
macondo OR « deepwater horizon » AND (« investigation report » OR « accident report »)	OU, ET, terme exact	All	193 000
		pdf	27 700
Google scholar			
deepwater horizon	Terme exact	All	41 400
deepwater horizon accident	Terme exact	All	23 200

Tableau 7 : résultats de différentes requêtes dans google et google scholar

On voit dans tous les cas que le nombre de résultats obtenus rend impossible un examen par un humain de chacun des documents. Même en ne prenant que les résultats de la première page de chaque requête, qui sont ceux considérés comme les plus pertinents par Google, le travail d’analyse serait considérable. Plutôt que de courir après une chimère d’exhaustivité, il va nous falloir fonctionner dans l’autre sens et choisir des sources de données dont on présume de l’intérêt pour continuer nos travaux de recherche sur l’accident de *Deepwater Horizon*.

Au début de nos travaux, nous avons commencé à dresser une cartographie, à la main, des sources accessibles par l’Internet pour la constitution de notre cas que nous appelons « modèle stratosphérique » :

Figure 16 : le « modèle stratosphérique »



Cette carte se lit de gauche à droite, c'est une arborescence du général au particulier. C'est une typologie de la structure de l'information à propos du cas *Deepwater Horizon* présente sur l'Internet. La dernière branche amenant à un site web unique. Cela permet de se rendre compte de la profondeur de recherche nécessaire pour l'obtention d'une information. A cela, nous ajoutons des liens transverses (en pointillés) entre différentes « feuilles ». Ces liens sont constitués au fur et à mesure de notre analyse des sources et nous cherchons à établir des rapports entre elles qui soient autres que des liens structurels, mais uniquement des liens *ad hoc*, créés pour le cas *Deepwater Horizon*. Par exemple, cela peut être un lien entre deux organisations pour la mise à disposition d'une information particulière, ou la création d'un site web collectif pour le partage de l'information. Il est très difficile de tenir à jour une telle cartographie, pas tellement pour les liens structurels (et encore, il est évidemment probable qu'il y ait bien d'autres sources d'informations disponibles qui n'apparaissent pas sur cette carte, mais cela reste finalement un travail de recensement), mais bien pour les liens spécifiques qui demandent un travail considérable d'évaluation.

1.2.2.2 L'évaluation de la donnée

Dès le début de nos travaux de recherche s'est posée la question de savoir en quoi une donnée était de qualité suffisante pour être traitée afin de construire notre représentation de l'accident. Comme nous l'avons vu, les sources sont multiples, provenant de différents organismes, retravaillées par les médias, critiquées par les « experts », commentées par le public et ainsi de suite… Notre travail de recherche dépend pourtant de la qualité et de la pertinence de ces données.

La qualité nous permet de répondre à la question de la possibilité d'utiliser une source de données tandis que la pertinence nous permet d'évaluer l'intérêt d'utiliser une source de données en rapport avec la recherche menée.

Assez généralement, nous pouvons considérer qu'avec l'accès via le Web, le « rapport signal sur bruit » s'est détérioré, mais d'un autre côté, on peut aussi travailler avec des données autrefois autrement plus difficiles d'accès comme le corpus de preuves d'un procès ou échanger par communications électroniques avec des personnes concernées par le sujet. Tout l'enjeu est donc de se situer dans ce « magma informationnel ».

Nous considérons que notre connaissance experte du domaine industriel dans lequel survient l'évènement *Deepwater Horizon* est un prisme nécessaire et *a priori* suffisant pour nous permettre d'effectuer cette évaluation et de distinguer le bon grain de l'ivraie, c'est-à-dire de déterminer les sources de données de meilleure qualité et qui nous semblent les plus pertinentes pour la constitution du cas *Deepwater Horizon*.

En pratique, nous reprenons l'approche de vérification présentée par Patton (1999). L'auteur propose un recoupement des informations par triangulation. Cette triangulation s'effectue sur deux plans : celui des approches (quantitative et qualitative) et celui des sources. A partir des résultats obtenus, un faisceau de preuves commence à se former, des informations convergent et permettent d'esquisser une réponse face à une interrogation. Une trame plausible de l'évènement se dessine et tend

à s'affirmer. Il ne s'agit pas d'une « quête de vérité », mais simplement de la capacité à identifier, tracer, suivre et affirmer une donnée pour en faire une information de qualité et pertinente pour nos travaux en ce qu'elle tend à produire un consensus social notamment parmi des experts. Nous partons donc à la recherche de sources de données susceptibles de produire de l'information d'intérêt. En particulier, il nous est apparu que le type de donnée la plus à même de satisfaire notre besoin de qualité et de pertinence correspond aux données d'un rapport d'enquête élaboré à la suite d'un accident ; par conséquent, nous allons maintenant nous pencher en premier lieu sur ce type de document.

1.2.3 Les rapports d'enquêtes

Concernant le cas *Deepwater Horizon*, nous avons identifié presque une trentaine de rapports d'enquêtes et d'autres types de document susceptibles de nous intéresser. Avant de rentrer en détail dans ces documents, nous présentons rapidement les limites conceptuelles et méthodologiques du rapport d'enquête et de son utilisation comme matériau de recherche.

1.2.3.1 Le processus d'investigation et la construction de l'accident

C'est l'enquête qui est le révélateur de l'accident comme objet d'étude : l'investigation façonne son objet et non l'inverse. L'enquête accident est menée selon des méthodes d'investigation diverses mises à la disposition des enquêteurs pour déterminer les causes de la survenance de l'évènement redouté (Sklet, 2004). En utilisant des méthodes d'enquêtes, les enquêteurs génèrent des données qui vont être structurées pour aboutir à un rapport d'enquête, document de dissémination des connaissances du cas d'accident en question.

Gephart (1984, 1992, 1993, 1997) et Gephart et al. (1990; 2010) ont montré que les rapports d'enquête remplissent un rôle social à travers leurs contenus techniques et que les conclusions dépendent largement des hypothèses de leurs auteurs sur le monde physique et social. Gephart rappelle cependant l'impérieuse nécessité de comprendre l'accident pour procéder à l'ajustement culturel, c'est-à-dire à la réaffirmation de l'autorité de l'Etat et de l'adhésion au modèle de société de l'ensemble des parties prenantes, notamment de la population (Gephart, Steier and Lawrence, 1990, pp. 28–29). Il s'agit de faire « comme si de rien n'était » tout en ayant accepté (et en faisant accepter) qu'un tel cas ne se reproduira pas. L'enquête est, selon Emerson (1981), « *la cérémonie en dernier ressort* ». Les arbitrages judiciaires effectués au pénal comme au civil relèvent en partie de cette logique et se nourrissent des rapports d'investigation techniques.

Au plan formel, les travaux de Benner Jr. (1975, 1985, 1989) ont été parmi les premiers à se pencher sur la méthode d'élaboration d'un rapport d'enquête. L'auteur *via* l'*IPRR*[17] propose une méthode de contrôle qualité de la rédaction des rapports d'enquêtes[18]. Dans la continuité du questionnement du rapport d'enquête, Burns (2000, para. 1.1.7 Weaknesses of Accident Reports) recense les faiblesses structurelles

17 Source : http://www.iprr.org/
18 Source : http://www.iprr.org/lib/qcp01.html

et méthodologiques dans l'élaboration du rapport d'enquête et propose comme solution une méthode formelle pour y remédier. Et plus récemment, Leveson (2011) rappelle, à propos de la subjectivité des données et des choix effectués pour « s'arrêter au bon moment » dans la recherche des causes :

> « *Une ultime raison pour laquelle une « cause profonde » peut être choisie est qu'elle est politiquement acceptable comme cause identifiée. D'autres évènements ou explications peuvent être exclus ou ne pas être examinés en profondeur parce qu'ils soulèvent des questions embarrassantes pour une organisation ou ses sous-traitants ou sont inacceptables sur le plan politique.* »
> (Notre traduction, Leveson, 2011, p. 20)

Pourtant, les conclusions tirées de ces rapports d'enquêtes peuvent servir d'orientations à visée opérationnelle ou politique et ces dernières peuvent donc considérablement modifier le paysage règlementaire ou normatif de la société. Ce constat doit impérativement être pris en considération dès lors que le rapport d'enquête devient matériau de recherche. Voici un schéma synthétique qui situe les données d'enquêtes dans leur processus d'élaboration :

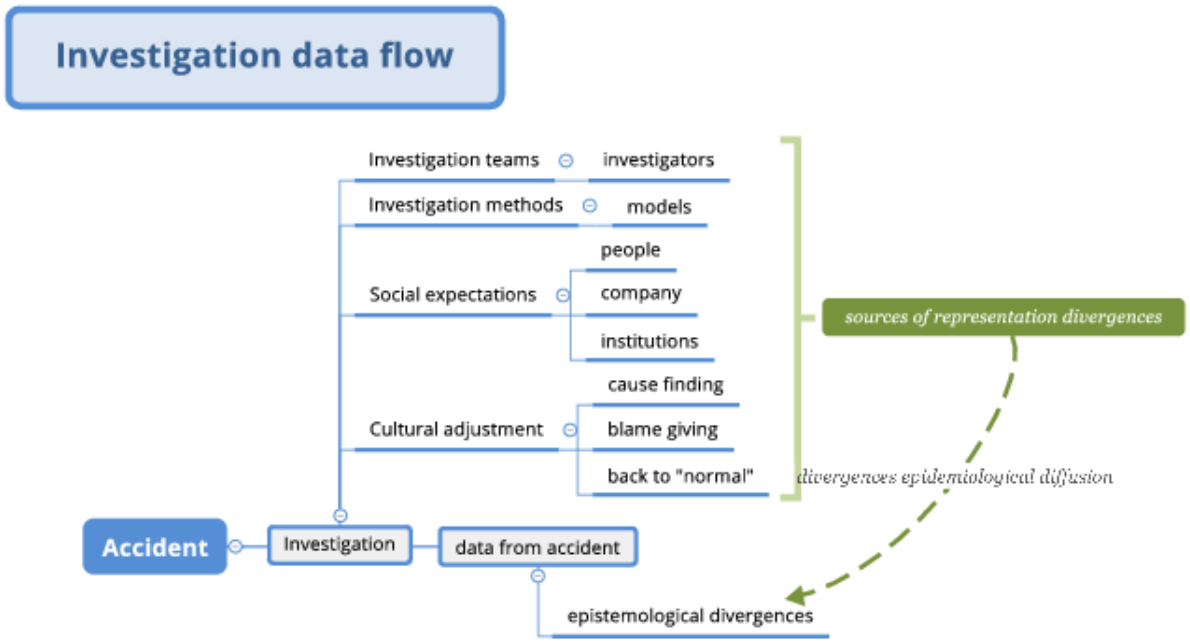


Figure 17 : le flux des données d'enquêtes

Le rapport d'enquête n'est pas un document à l'objectivité parfaite, mais il est celui dont la démarche d'élaboration nous paraît la plus intègre et transparente. Or, selon notre perspective, il ne s'agit pas de prendre les conclusions ou les données des rapports pour « vraies », mais de les comparer à d'autres données dont il est possible de tracer le processus d'élaboration. A cet égard, les investigations techniques officielles produisent les données les plus abouties à la suite d'un accident et nous allons donc nous appuyer dessus pour la suite de nos travaux[19].

19 A ceux-là, nous ajoutons aussi l'ensemble des documents préliminaires à leur élaboration que nous avons pu consulter.

Ensuite, nous avons également identifié d'autres sources qui nous paraissent fiables et pertinentes pour l'étude de cas, notoirement, les constatations de fait et conclusions de droit issues du procès opposant BP et les États-Unis, contentieux le plus important à ce jour de toute l'histoire de ce pays. La liste de ces sources que nous considérons comme étant les plus solides est en annexe de ces travaux. Nous proposons une rapide analyse quantitative de ce corpus de référence[20].

1.2.3.2 Quelques chiffres sur les rapports d'investigation sur Deepwater Horizon

Nous recensons quarante-deux documents de référence (hors sites web) qui totalisent plus de quatre-mille-trois-cents pages. Les rapports d'enquête sont majoritaires en nombre et en nombre de pages, avec vingt-six documents qui totalisent plus de trois-mille-six-cents pages.

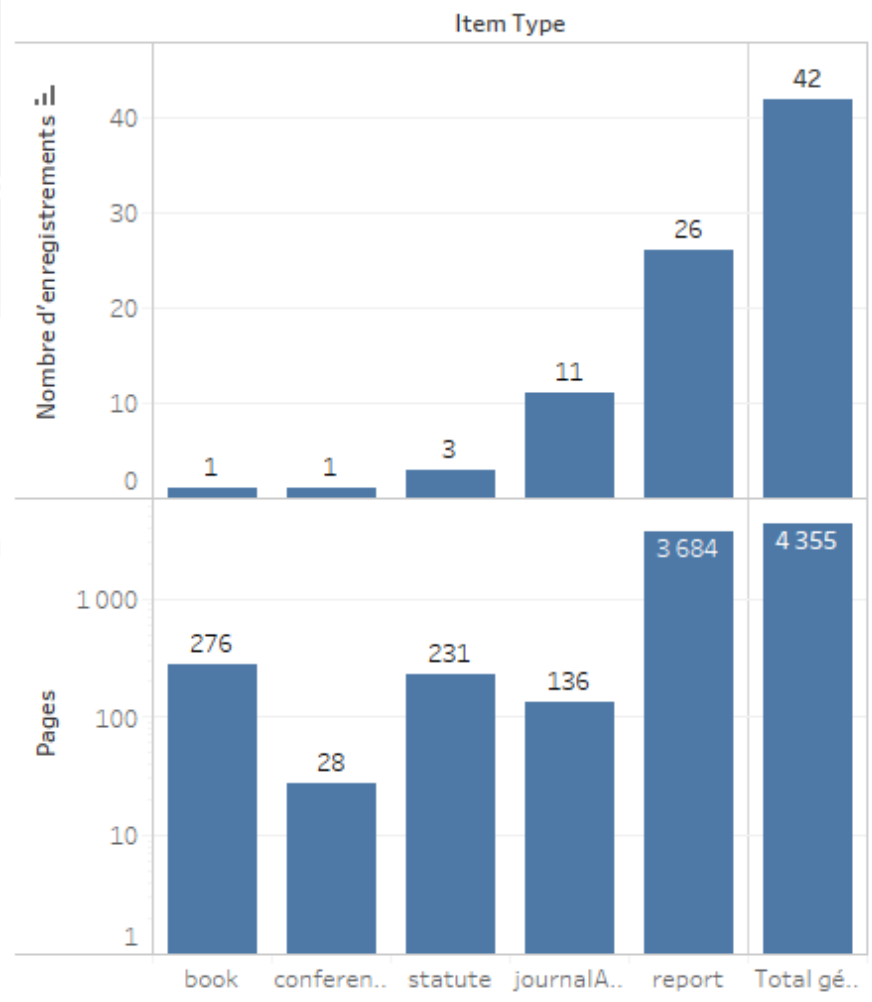


Figure 18 : les documents de référence

20 L'analyse a été effectuée en utilisant le logiciel Tableau Public (*Tableau Public*, no date).

Si l'on regarde maintenant chaque type de document en fonction de sa date de publication, on obtient les graphiques suivants :

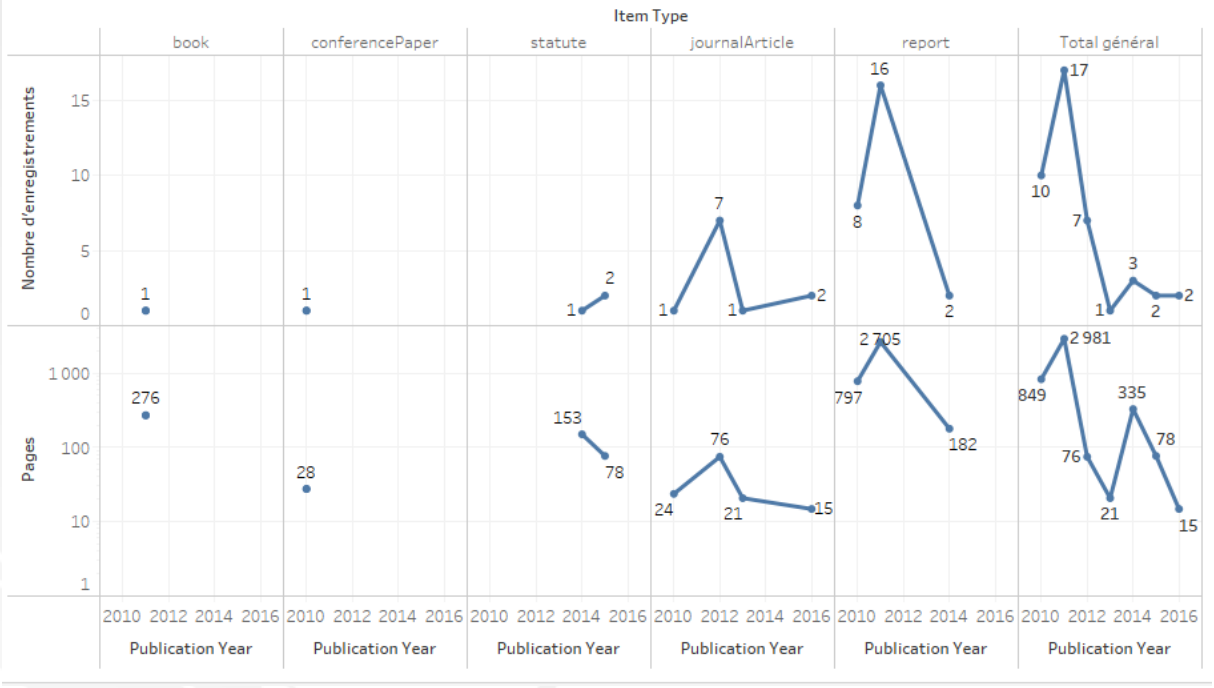


Figure 19 : évolution chronologique de la production du corpus depuis 2010

La production des documents est la plus importante entre 2010 et 2012 et atteint un pic en 2011. Cette production est dans l'ensemble liée aux publications des rapports d'enquêtes pendant cette période. Suite à cela, nombre d'articles de science dès 2012 (dans leur grande majorité) puis des documents juridiques (*statute*) dès 2014 sont publiés. De façon très corrélée, la plus grande quantité de pages produites est liée la publication des rapports d'enquêtes. A noter que la publication des conclusions de droit (*statute*) suite au procès produisent trois-cent-trente-cinq pages au total, mais surtout, font appel à de très nombreux autres documents dont le recensement est disponible sur le site qui archive l'ensemble des documents produits pendant le procès[21].

Nous revenons aux rapports et documents d'enquêtes, qui ont servi de sources aux autres travaux scientifiques et juridiques. Nous avons créé de façon tout à fait empirique une petite cinquantaine d'étiquettes qui nous a parue pertinente dans le but de qualifier ces rapports[22]. Cela nous permet de proposer une vue synthétique de ces documents en fonction de leur contenu. L'ensemble des données est manipulable à l'adresse https://public.tableau.com/profile/bobby9057#!/ et nous proposons ici

21 Il y a un index pour chaque phase du procès (*MDL 2179 Trial Docs - Phase One*, no date; *MDL 2179 Trial Docs - Phase Three*, no date; *MDL 2179 Trial Docs - Phase Two*, no date).

22 Les étiquettes et leurs attributions sont tout à fait sujettes à discussion : accident investigation, administrative/government, BOP, calculation/modeling, causes of the blowout, challenge trade-offs, change, command, complexity, containment, contractor, decision-making, dispersant, drilling rig, effort, emergency operations, energy policy, environment, failure, flag state, flow rate/oil budget, forensic, funding/investment, geological reservoir, human factor, lessons learnt, management, marine casualty, nation/region wide, new/novelty, offshore reform, oil fate, oil spill, operation/engineering, organization, recommendations, response, responsible party, restoration, safety (generic), shut-in, SONS, well, well control.

quelques extraits. Les rapports dont le sujet principal est la détermination des causes de la survenance du *blowout* sur la plateforme sont les suivants :

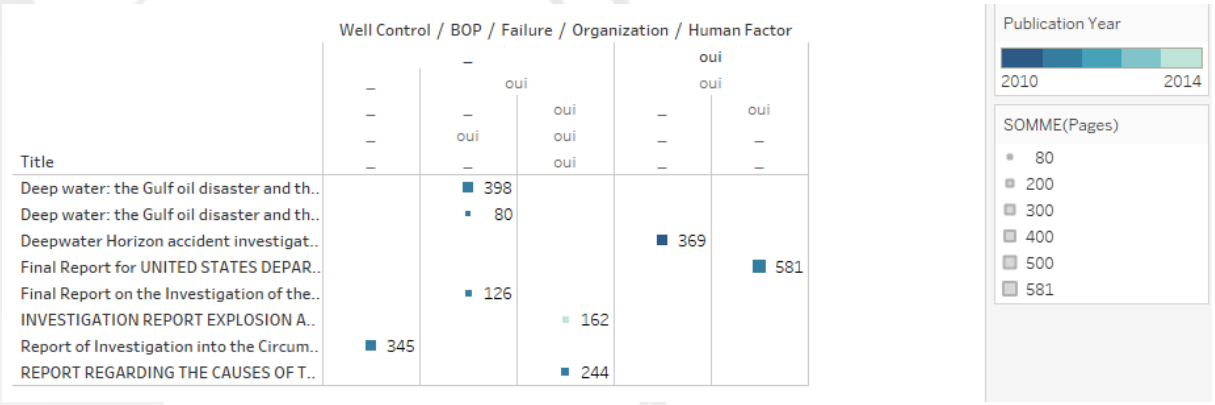


Tableau 8 : les rapports d'enquêtes sur la survenance du blowout

C'est un tableau croisé qui permet d'associer, pour chaque rapport dont le sujet principal est la détermination des causes, un ou plusieurs facteurs (*Well control*, *BOP*...) apparus comme déterminants dans la survenance du *blowout*. L'ordre des étiquettes correspond à l'ordre en colonne (à la manière de portes logiques). De la même façon, on peut visualiser certains rapports qui traitent de sujets très spécifiques comme l'usage de dispersant :

Title	Oil Spill / Dispersant oui oui
Response and clean-up technology research and development an..	28
Revisions Needed to National Contingency Plan Based on Deepw..	42

Tableau 9 : les rapports qui traitent particulièrement de l'usage de dispersant

Enfin, si l'on s'intéresse à l'ingénierie déployée et aux solutions proposées pour tenter de fermer le puits, on pourra consulter les rapports rassemblés dans le tableau en page suivante :

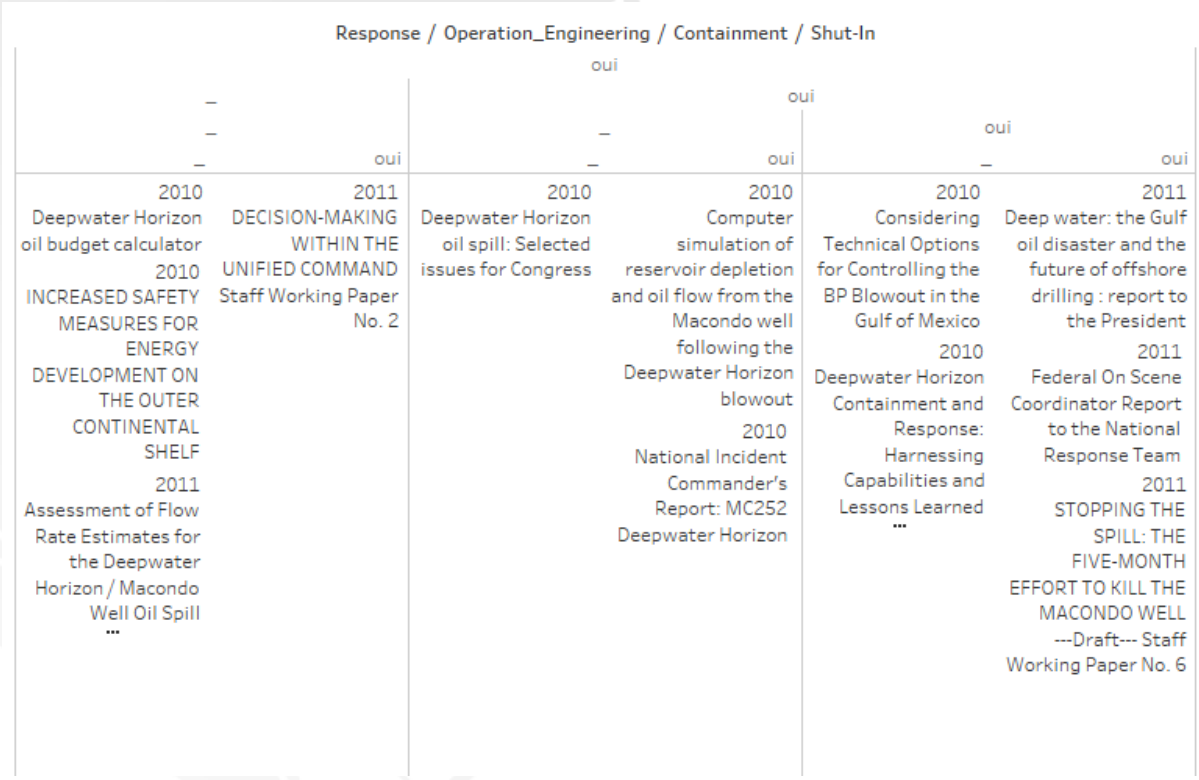


Tableau 10 : les rapports d'enquêtes sur l'intervention et certaines spécificités

A cela nous pouvons adjoindre trois interviews que nous avons conduites en mars 2016, avec « ceux qui ont fait *Deepwater Horizon* » : David Westerholm, directeur du bureau *Response and Restoration* de la *NOAA*[23] ; le Pr. Steven Chu, prix Nobel de physique, Professeur à l'université de Stanford, ancien secrétaire d'État à l'énergie de l'administration Obama[24] ; et le Dr. Paul Hsieh, géologue hydrologue, à *l'USGS*[25].

1.3 Du forage des données relatives à *Deepwater Horizon*

Dans cette section, nous plongeons « *in media res* » (Latour dans Dumez, 2016), pour mieux appréhender le cas *Deepwater Horizon* (et la gestion du déversement d'hydrocarbures dans l'environnement qui a suivi) en tant qu'objet d'étude. Nous abordons la problématique du passage de données à une « connaissance » relative à un accident. Dans un premier temps, nous nous intéressons à la manière dont la science s'est saisie du cas, puis nous développons l'exemple de la détermination du paramètre « *oil budget* », dont la mesure permet une estimation du coût de la catastrophe et qui a servi de jauge de son ampleur. Enfin, nous posons les orientations théoriques qui ont guidé la présente recherche vers une méthode systématique de structuration des connaissances à propos d'un accident.

23 Source : https://www.noaa.gov/
24 Source : https://www.energy.gov/
25 Source : https://www.usgs.gov/

1.3.1 Le traitement de *Deepwater Horizon* par la science

Nous portons ici un regard critique sur l'utilisation par la communauté scientifique des données issues du cas *Deepwater Horizon*. Nous ne pouvons lire l'entièreté des publications consacrées à cet évènement : c'est le problème épistémologique fondamental de cette thèse en matière de recherche et de consultation d'informations**.** A l'heure de l'Internet et des bases de données scientifiques accessibles en ligne, il nous paraît impossible d'établir un état de l'art pertinent sans passer par des intermédiaires algorithmiques pour nous faciliter l'analyse.

Nous devons cibler d'ores et déjà les publications qui *a priori* traitent de *Deepwater Horizon* en tant qu'« accident » pour constituer un corpus que nous supposerons représentatif de la science produite à partir du cas. Notons que si notre définition de « l'accident », proposée en section 1, nous permet de mieux repérer les différents travaux, elle ne correspond pas à celle utilisée dans le champ des *safety studies*. Nous devons donc également examiner ce que les experts et les chercheurs entendent par « accident » en tant qu'objet d'étude.

Dans la base de données *ScienceDirect*[26] nous effectuons la requête suivante : « *accident* » AND (« *Deepwater Horizon* » OR « *Macondo* ») et nous obtenons 854 résultats. Nous souhaitons savoir, dans ce corpus de documents, sous quel angle l'accident de *Deepwater Horizon* est abordé par la recherche. Nous utilisons le logiciel *KH coder*[27]*,* dédié à la fouille de texte, *open source* et développé par le Professeur Higuchi[28]. L'ensemble des références, notamment la description des algorithmes utilisés, est disponible sur le forum du site[29]. Nous précisons que l'algorithme de lemmatisation[30] (*NLP POS Tagger*), est celui créé par Stanford. *KH coder* est utilisé dans de nombreux travaux de recherche notamment en sociologie[31]. Il s'agit donc ici de fouiller l'ensemble des publications scientifiques ciblées pour obtenir une « image » et des tendances quant aux perspectives selon lesquelles la science ausculte le cas *Deepwater Horizon*. Nous avons choisi par commodité (et économie de temps-machine) de ne travailler que sur les métadonnées - le titre, les mots-clés et le résumé de chaque article – en postulant leur représentativité. Puis, nous avons effectué une recherche par mots-clés en contexte, qui permet de situer dans son contexte l'emploi de mots-cibles, ou plutôt du lemme de chaque mot (ou de combinaisons de mots). La méthodologie utilisée et les références pour les calculs statistiques sont décrites dans (A.5.6 [KWIC Concordance] Koichi HIGUCHI, 2016). Brièvement, le logiciel compte, dans une étendue maximale de cinq mots à droite et à gauche du mot-cible, l'ensemble des mots qui apparaissent et, en fonction de leur position et de l'algorithme utilisé, propose un classement. La création d'un réseau de co-occurrence par associations de mots est également intéressante. De nombreux algorithmes de classification – *scoring* – sont disponibles (A.5.11 [Co-Occurrence Network] of words

26 Source : https://www.sciencedirect.com/
27 Source : Source : http://khcoder.net/en/index.html
28 Source : http://koichi.nihon.to/psnl/en/
29 Source : (*KH Coder / Discussion / Open Discussion:Co-occurrence Networks*, no date b).
30 Nous allons revenir en détail sur ce procédé d'analyse textuelle dans le chapitre 3.
31 Source : http://khcoder.net/en/bib.html?year=all&lang=English&key=

Koichi HIGUCHI, 2016) et apportent différents points de vue sur les résultats. Nous renvoyons le lecteur en annexe de cette thèse pour le traitement des données par *KH coder* ; nous ne présentons ici que nos résultats. Nous précisons également que l'interprétation des résultats nécessite une expertise du domaine considéré.

1.3.1.1 Les perspectives de la science sur l'accident de Deepwater Horizon

Nous classons les résultats avec l'algorithme Z-score. Ce dernier calcule l'écart par rapport à une moyenne dans un ensemble de variables centré-réduit. Il permet un classement relatif des variables (des unités lexicales) entre elles en fonction de leur écart à la moyenne ; cela signifie ici que plus le score est élevé, plus la variable (l'unité lexicale) est utilisée en association avec le mot-cible par rapport à un autre couple. Cet algorithme élimine également les mots utilisés très fréquemment associés avec le mot-cible s'ils sont aussi utilisés de manière fréquente avec d'autres termes. C'est donc un indicateur assez discriminant pour distinguer des couples de mots (des associations de deux mots-cible) qui caractérisent l'emploi du mot-cible. Dans les résultats qui vont suivre, nous ne montrons, sauf mention contraire, que les dix résultats les plus significatifs[32]. Le mot-cible est « *accident* ».

Il y a quatre-cent-quatre-vingt-sept occurrences dans notre corpus et la valeur du Z-score figure en dernière colonne du tableau suivant :

Word	POS tag	TT	TL	TR	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	Z-score
major	Adj	103	98	5	1	2	0	11	84	0	0	3	0	2	49,94
precursor	Noun	9	1	8	0	1	0	0	0	8	0	0	0	0	16,81
investigation	Noun	19	3	16	1	0	1	1	0	11	1	3	0	1	16,64
causation	Noun	7	0	7	0	0	0	0	0	7	0	0	0	0	16,09
prevention	Noun	15	3	12	1	0	1	1	0	11	1	0	0	0	15,22
FUKUSHIMA	Proper Noun	10	5	5	0	0	0	1	4	0	2	0	2	1	13,39
fatal	Adj	4	4	0	0	1	2	0	1	0	0	0	0	0	13,04
frequent	Adj	7	6	1	0	1	2	0	3	0	0	0	1	0	12,96
maintenance-related	Adj	3	2	1	0	0	0	2	0	0	1	0	0	0	12,69
occur	Verb	21	1	20	0	1	0	0	0	6	6	4	2	2	12,09

Tableau 11 : classement Z-score des unités lexicales associées à « *accident* »

L'accident est considéré « majeur », adjectif qui permet de faire le lien avec les accidents industriels et probablement de mettre de côté d'autres catégories d'accident (travail, circulation, domestique…). L'unité lexicale « *major* » est nettement plus utilisée avec « *accident* » que « *occur* ». On trouve des termes qui sont liés à l'enquête avec « *precursor* », « *investigation* » ou « *causation* ». Le nom propre Fukushima apparaît, faisant probablement référence à l'accident nucléaire de Fukushima Daiichi (2011) ; il y a donc des articles qui associent cet accident à celui de *Deepwater Horizon*. Les autres adjectifs (« *fatal* », « *frequent* », « *maintenance-related* ») qualifient l'accident. On observe la dichotomie usuelle des matrices de risque entre la

32 L'ensemble de chaque liste est disponible en annexe.

gravité et la probabilité, qui positionnent l'accident dans le corpus. D'autre part, nous avons constaté empiriquement que les accidents sont souvent associés à un nom propre qui le désigne et ce nom propre en fait une référence unique. Nous avons voulu voir les résultats obtenus en recherchant uniquement les noms propres entourant le mot-cible « *accident* » :

Rank	Word	TT	TL	TR	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	Z-score
1	FUKUSHIMA	10	5	5	0	0	0	1	4	0	2	0	2	1	15,15
3	Krimsk	1	1	0	0	0	0	0	1	0	0	0	0	0	11,71
9	Chernobyl	2	2	0	0	0	0	1	1	0	0	0	0	0	10,36
10	Horizon	18	15	3	0	2	1	0	12	0	0	0	2	1	8,78
11	Deepwater	18	15	3	2	1	0	12	0	0	0	2	1	0	8,59
12	Columbia	1	1	0	0	0	0	0	1	0	0	0	0	0	8,22
19	Penglai	2	2	0	0	1	0	1	0	0	0	0	0	0	6,66
22	Alpha	3	2	1	0	0	0	0	2	0	0	0	1	0	5,99
23	Piper	3	2	1	0	0	0	2	0	0	0	1	0	0	5,99
31	Hebei	1	1	0	0	0	0	1	0	0	0	0	0	0	3,27
32	Spirit	1	1	0	0	0	0	0	1	0	0	0	0	0	3,11
38	Texas	1	0	1	0	0	0	0	0	0	1	0	0	0	1,99
43	Macondo	2	2	0	1	0	0	0	1	0	0	0	0	0	0,69
44	BP	8	7	1	0	6	0	0	1	0	0	0	1	0	0,41
52	Challenger	1	1	0	0	0	1	0	0	0	0	0	0	0	- 0,08
61	Dai-ichi	2	0	2	0	0	0	0	0	0	0	0	0	2	- 0,12
63	Mile	1	0	1	0	0	0	0	0	0	0	1	0	0	- 0,12
66	Soviet	1	0	1	0	0	0	0	0	0	0	1	0	0	- 0,12
69	Japan	1	0	1	0	0	0	0	0	0	0	0	0	1	- 0,15
70	Mumbai	1	0	1	0	0	0	0	0	0	0	1	0	0	- 0,15
74	Daiichi	2	0	2	0	0	0	0	0	0	0	2	0	0	- 0,17
77	Buncefield	1	0	1	0	0	0	0	0	0	0	1	0	0	- 0,19
81	Dalian	4	4	0	1	0	3	0	0	0	0	0	0	0	- 0,21
88	Island	1	0	1	0	0	0	0	0	0	0	0	1	0	- 0,27

Tableau 12 : classement Z-score avec noms propres associés à « *accident* »

Après un recoupement rapide, on s'aperçoit que dans le corpus considéré, il y a un nombre conséquent d'accidents industriels apparaissant dans des articles où il est question en premier lieu de l'accident de *Deepwater Horizon*. Nous supposons qu'il s'agit de comparaisons, pour situer l'accident de *Deepwater Horizon* dans le retour d'expérience et dans la mémoire.

Si nous nous intéressons à d'éventuelles considérations sur les modèles (unité lexicale « *model* »), nous trouvons quatre-cent-trente-et-une occurrences dans le corpus.

En ciblant les adjectifs qui qualifient le modèle on obtient :

	Word	TT	TL	TR	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	Z-score
1	predictive	10	6	4	0	2	0	1	3	3	1	0	0	0	17,17
2	mental	6	6	0	0	0	0	0	6	0	0	0	0	0	14,64
3	swiss	4	4	0	0	0	0	4	0	0	0	0	0	0	12,64
4	numerical	7	4	3	1	0	0	1	2	0	1	0	1	1	12,18
5	linear	5	2	3	0	0	1	0	1	0	0	1	1	1	12,14
6	epidemiological	3	2	1	0	0	0	0	2	0	0	0	1	0	11,64
7	barrier-based	2	2	0	0	0	0	2	0	0	0	0	0	0	11,03
8	multi-plant	2	2	0	0	0	0	2	0	0	0	0	0	0	11,03
9	probabilistic	8	8	0	1	0	3	1	3	0	0	0	0	0	10,69
10	hybrid	3	3	0	0	0	1	0	2	0	0	0	0	0	9,40

Tableau 13 : classement Z-score adjectifs associés à « *model* »

On constate une forte prégnance de ce qui semble être les attributs du modèle de Reason (2013) (« *swiss* », « *epidemiological* », « *barrier-based* »), ce qui est d'ailleurs conforme à son usage très répandu dans les *safety studies* et l'industrie. Les termes « *predictive* », « *numerical* » et « *probabilistic* » peuvent être les marqueurs d'une approche quantitative de l'accident. En soi, il y a de nombreux attributs avec des valeurs élevées qui laissent à penser que la modélisation du cas *Deepwater Horizon* a son importance en science.

On notera, dans le bas du classement, l'unité lexicale « *organizational* » :

221	organizational	2	1	1	0	1	0	0	0	0	1	0	0	0	0,1

Tableau 14 : unité lexicale « organizational » avec « model »

En étant positionnée quasiment sur la valeur moyenne de l'algorithme, l'association de cette unité lexicale avec « *model* » n'apporte pas de signification particulière, soulignant l'équivalence entre « modèle » et « modèle organisationnel ».

Au final, l'intérêt premier de la science reste la détermination des causes et la volonté de pouvoir qualifier et quantifier l'accident. Les modèles d'accidents sont présents dans le corpus et de nombreux descripteurs de la nature de ces modèles nous apportent qu'il semble exister deux branches de la modélisation qui cohabitent : l'une concerne l'approche facteurs humains et organisationnels, probablement autour du modèle de Reason, l'autre quantitative, basée sur des modèles calculatoires probabilistes. Nous allons montrer maintenant en détail la portée des travaux scientifiques à propos du cas *Deepwater Horizon*.

1.3.1.2 Le réseau de co-occurrence du mot-cible Deepwater Horizon

Nous cherchons maintenant à déterminer toutes les associations possibles autour de *Deepwater Horizon*. Nous allons construire un réseau de co-occurrence qui va représenter les associations de termes les plus fréquentes et en lien avec le mot-cible *Deepwater Horizon* (ET booléen entre les deux termes) en page suivante :

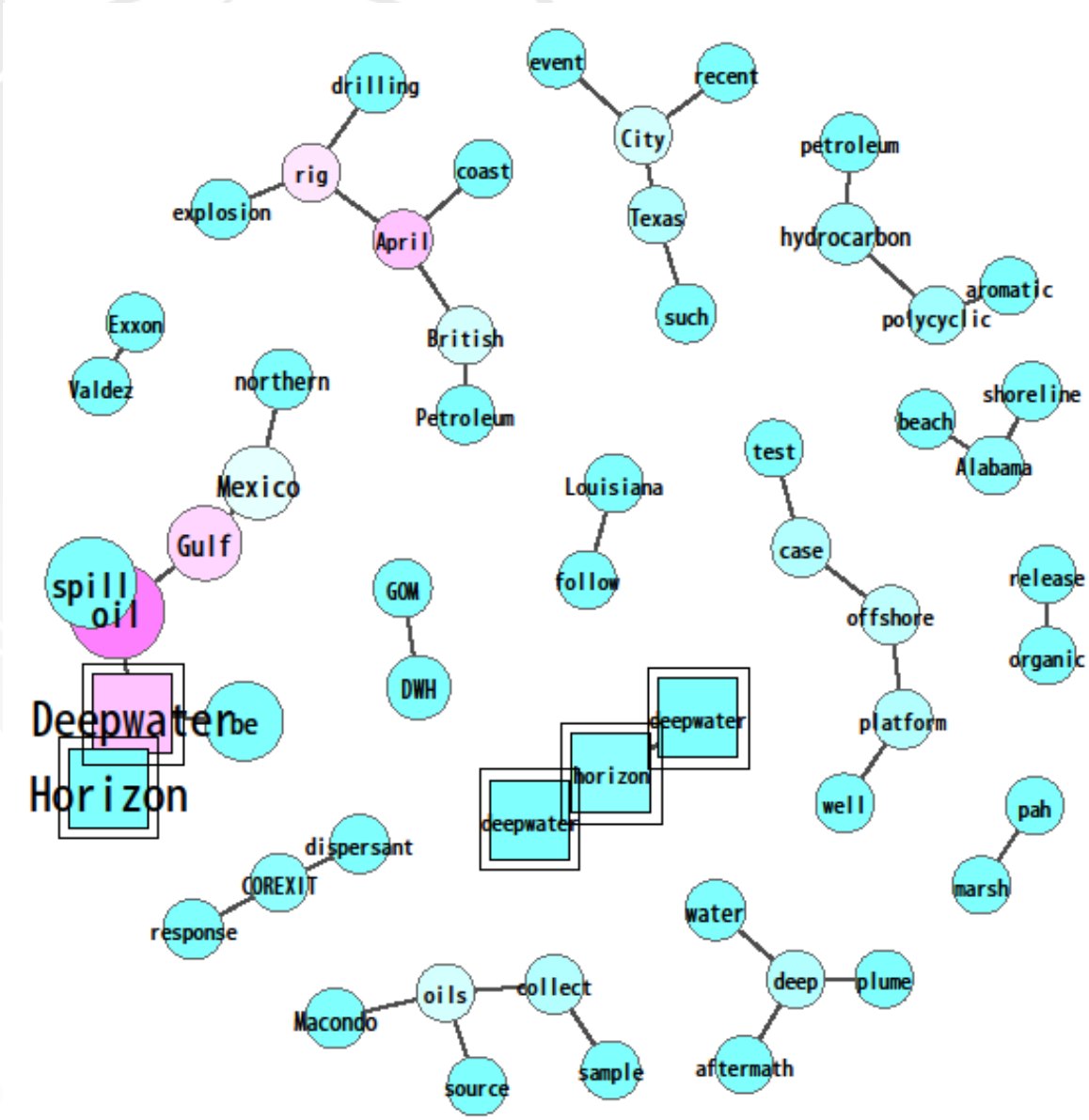


Figure 20 : la science sur le cas *Deepwater Horizon*

La figure carrée correspond au(x) mot(s)-cible(s). Le lien de co-occurrence est porté par la ligne entre deux cercles. Par souci de lisibilité, seul le lien le plus fort entre deux termes est représenté ici et l'épaisseur de la ligne correspond à sa force. Ce n'est donc pas la répartition spatiale qui exprime la force de l'association. C'est l'indice de Jaccard qui est utilisé pour calculer la force de la co-occurrence :

$$Jaccard_w = \frac{a}{F_1 + F_2 - a} \qquad \text{(équation 5)}$$

Avec a la fréquence du mot w avant et après le mot-cible dans l'étendue considérée ; F_1, la fréquence du mot-cible dans l'ensemble du corpus et F_2, la fréquence du mot w dans l'ensemble du corpus considéré. Enfin, la taille de chaque cercle est proportionnelle à la fréquence de l'unité lexicale (le nombre d'apparitions) dans le corpus.

Avec ce graphe, il est possible de proposer une représentation du cas *Deepwater Horizon* telle qu'il est considéré dans le corpus, même si un degré d'expertise du domaine reste nécessaire pour l'interpréter.

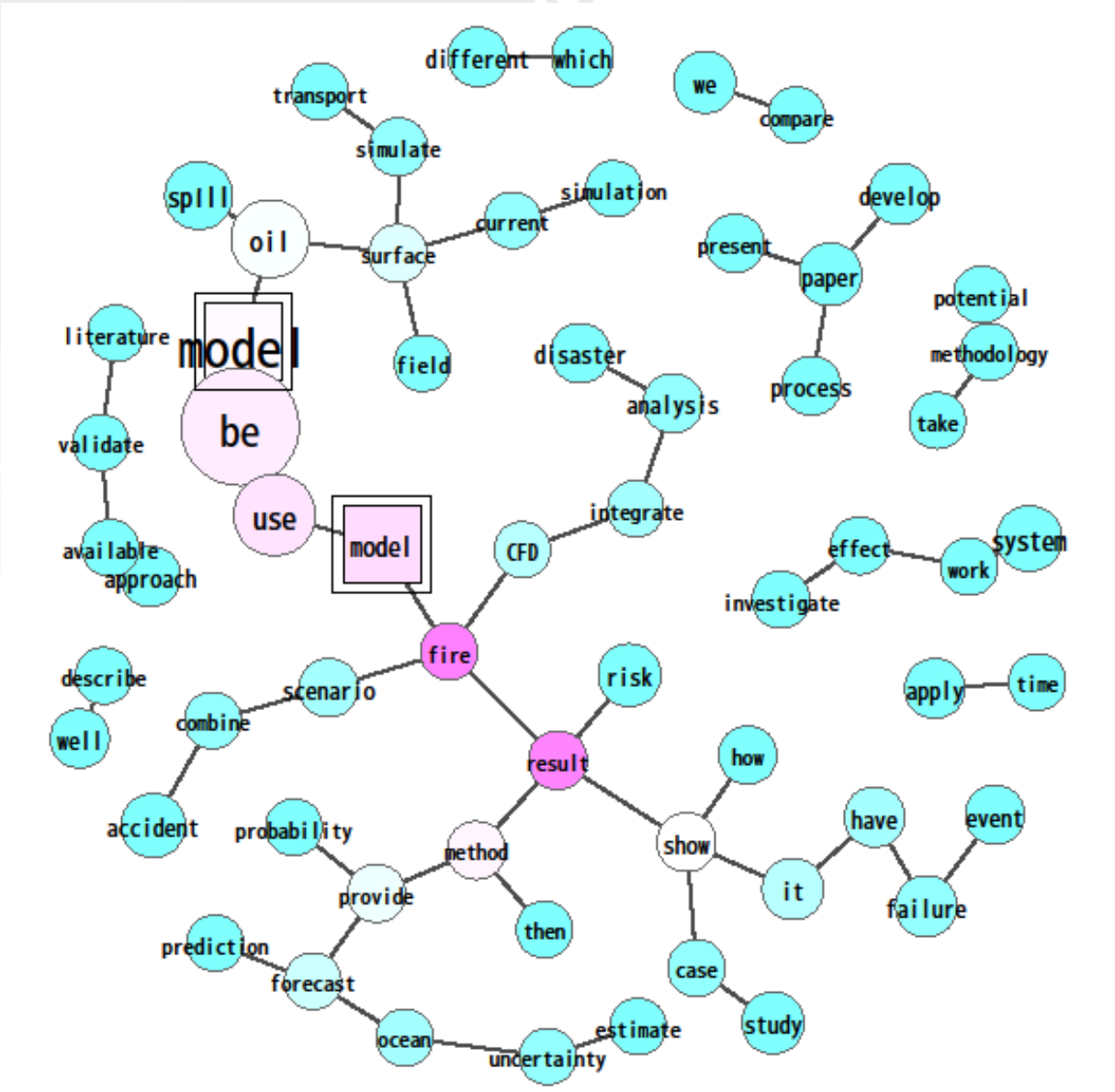


Figure 21 : la considération du cas *Deepwater Horizon* par la modélisation

On peut dire que :

- il s'agit d'un déversement accidentel d'hydrocarbures en mer, plus précisément dans le nord du Golfe du Mexique ;
- il y a eu une explosion sur un *rig* de forage *offshore* en avril. Il y a un lien avec BP ;
- il y a eu des tests effectués sur un puits ;
- il y est fait référence à deux accidents précédents, le naufrage du supertanker Exxon Valdez, et l'explosion de la raffinerie de Texas City ;
- la qualité du produit déversé fait l'objet d'un important travail de description physico-chimique ;
- l'intervention a nécessité l'utilisation de dispersant, le COREXIT ;

- il y a un lien entre les conséquences (du déversement), sa forme (le panache) et l'environnement (l'eau, la profondeur).

On peut encore apporter des précisions sur la représentation du cas *Deepwater Horizon* au travers de sa modélisation en cherchant le mot-cible « *model* ». Sur le graphe en page suivante, on rajoute l'illustration par code couleur de la centralité par intermédiarité, c'est-à-dire l'importance d'un nœud (ici une unité lexicale) dans le réseau par son rôle de mise en relation d'un nœud à un autre via le transfert d'information au sein du réseau.
On voit ici que la modélisation du cas *Deepwater Horizon* tourne autour de deux axes :

- l'incendie ;
- la nappe d'hydrocarbures.

En effet, les nœuds « *fire* » et « *oil* » sont ceux qui relient le plus fortement le reste du réseau au mot-cible « *model* ».

Après cette « revue » de littérature, il nous semble que le positionnement de la science sur le cas *Deepwater Horizon,* ramenée à une chronologie telle qu'exposée dans le rappel des faits, peut être résumée ainsi :

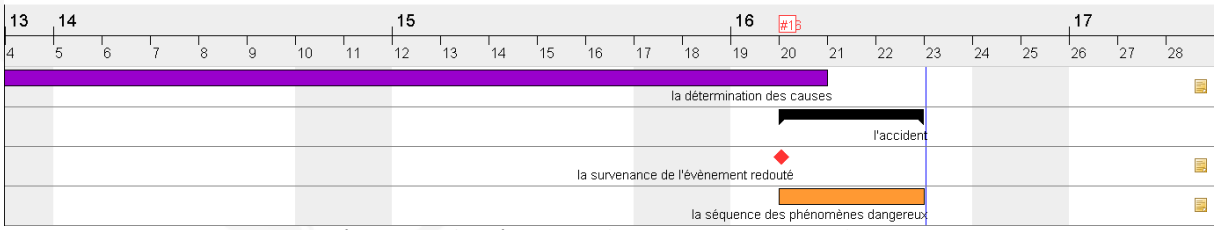


Figure 22 : la science sur le cas *Deepwater Horizon*

Soit :

- l'utilisation de modèles d'accidents pour la détermination des causes ;
- l'examen de l'évènement redouté et la séquence de phénomènes dangereux, l'explosion et l'incendie ;
- la modélisation du déversement d'hydrocarbures en mer, particulièrement la nappe de pétrole.

Or, il y a tout de même un consensus formel au sein des rapports d'enquêtes sur le fait qu'il y a eu une pollution majeure qui a été stoppée par l'action des hommes ; à notre connaissance, il n'y a aucun rapport ou papier de science qui affirme le contraire. Ce n'est ni l'épuisement du réservoir géologique ni l'atteinte de l'équilibre hydrostatique, mais bien le travail de la plus grande organisation d'intervention en temps de paix qui a permis de mettre fin à cette pollution. Cette période, que nous nommerons « l'intervention », a duré quatre-vingt-sept jours. Par conséquent, le « domaine d'existence anormal » dans lequel l'évènement *Deepwater Horizon* a projeté ses acteurs (cf. notre définition de « l'accident » en section 1.1.3), correspond à une activité intense de quatre-vingt-sept jours qui ne peut être détachée de « l'accident » tant elle contribue à définir le choc qu'il a constitué. D'autant plus qu'il y a, en sus de certains rapports d'enquêtes qui relatent cette séquence, des articles scientifiques (Camilli *et al.*, 2012), (Hickman *et al.*, 2012), (Lubchenco *et al.*, 2012) et (McNutt, Camilli, *et al.*, 2012; McNutt, Chu, *et al.*, 2012) dont les auteurs, nous y reviendrons grandement par

la suite, ont été acteurs de cette intervention et où leur travail a permis de mettre fin à l'accident. L'intervention est une réponse face à la « faillite de l'ingénierie » conduite en situation normale. L'accident de *Deepwater Horizon* se caractérise avant tout par la nature de l'évènement redouté : une fuite d'hydrocarbures dans l'environnement, avec l'hypothèse que les hydrocarbures sont des polluants et que leur déversement dans l'environnement est inacceptable dans nos sociétés (ce qui renvoie à l'usage d'« accident » dans la langue).

Le déversement est directement fonction du temps et la relation est même proportionnelle et il peut être modélisé comme un système intégrateur pur :

- le débit du puits, que l'on peut considérer comme une constante avec un *decay*[33] est représenté par une pente estimée à -0.1047[34] ;
- le volume relâché dans l'environnement, qui est l'intégrale de ce débit (somme cumulative par rapport au temps),

$$V = \int_0^t dqdt \; avec \; q = \frac{\rho v}{t}$$ soit (équation 6).

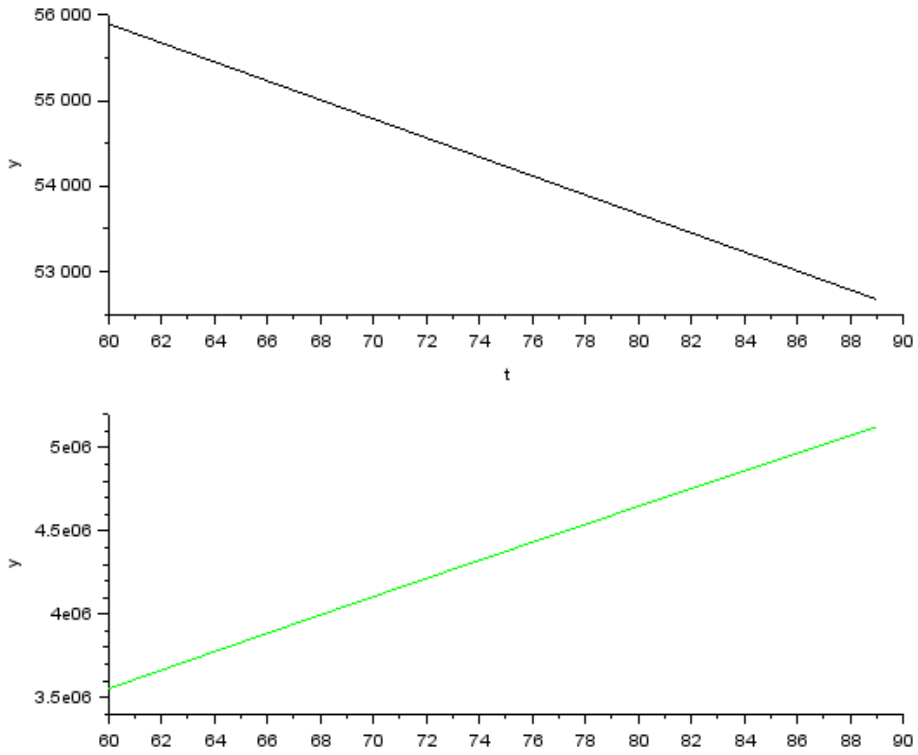


Figure 23 : la modélisation de l'accident de Deepwater Horizon

Ci-dessus, une illustration graphique de cette modélisation[35] : le graphique du haut est celui du débit du puits et le graphique du bas celui du volume relâché. La réponse à une rampe (ici le débit qui diminue au cours du temps) est une rampe (ici le volume relâché

33 Le *decay* représente la déplétion, l'épuisement du réservoir géologique, qui est fonction de la pression et donc qui fait que le débit du puits s'amenuise progressivement au cours du temps.
34 Estimation rapide effectuée sur la base du « *August model* » présenté par McNutt et al. (2012). Il s'agit d'une pente calculée par interpolation linéaire.
35 Les axes des ordonnées ne sont pas à la même échelle. L'axe des abscisses matérialise le temps (le pas est le jour), le 87ème jour correspond au 15 juillet 2010.

qui augmente). Le système est donc naturellement instable et la sortie (le volume relâché) tend virtuellement vers l'infini (dans le monde réel, avec la déplétion représentée par le *decay*, cela serait l'équilibre hydrostatique entre la pression de sortie du puits et la pression de l'eau à cette profondeur).

Nous venons de montrer que cette intervention n'a pas été considérée par la science dans son abord du cas *Deepwater Horizon,* comme si l'accident changeait de dynamique, passant du *blowout* à l'explosion puis à la pollution sans que jamais on ne s'interroge sur le travail et sur l'organisation humaine mis en œuvre pour y mettre un terme, alors même que « le domaine d'anormalité » grossissait avec le temps (cf. figure 1 *supra*, et section 1.1.1.1 sur notre définitin de l' « accident »). A notre sens, il ne s'agit pas d'un manquement , mais plutôt d'une inadéquation des théories explicatives de l'accident, l'inadaptation des modèles d'accident existants (Travadel and Guarnieri, 2016) et probablement une réflexion à pousser sur l'ontologie de l'accident. Le cas *Deepwater Horizon* nécessite une approche nouvelle pour enrichir la connaissance à partir des masses considérables de données disponibles.

1.3.2 L'attribut *oil budget*

Nous avions souligné en 1.2.2 *supra* la nécessité de valider les données retenues dans un tel océan ou, du moins, de les comparer et de conserver la trace des liens entre des conclusions et des données lorsque ces dernières sont contradictoires. Nous montrons ici la nature de la problématique, à travers l'un des premiers exemples auquel nous avons été confrontés, à savoir la détermination par les experts du paramètre « *oil budget* », c'est-à-dire le volume d'hydrocarbures relâché par le puits.
Il s'agit de la quantification d'une caractéristique physique qui a permis de donner du sens au phénomène « puits d'hydrocarbures sous-marin en éruption ». Le processus de « réalisation » de l'attribut *oil budget* est une illustration de la théorie de la traduction développée par Callon (1986). Un cortège de valeurs pour cet attribut a été développé en deux temps, d'abord pendant l'intervention puis pendant le procès à l'encontre de BP. Cette caractéristique a été fondamentale à la fois pendant l'intervention sur le puits et pour le dimensionnement des opérations de récupération et de nettoyage des hydrocarbures, mais aussi, et surtout, dans le calcul du montant de l'amende et des dommages et intérêts fixés par la justice américaine.

1.3.2.1 Le processus pendant l'intervention

Un document de l'administration fédérale (The Federal Interagency Solutions Group, *et al.*, 2010) propose un calculateur du volume total d'hydrocarbures relâché par le puits comme un outil d'aide à la décision pendant l'intervention. En effet, un document administratif (*ICS Form 209*) est créé dans le cadre de la gestion d'un déversement accidentel d'hydrocarbures ; mais *Deepwater Horizon* étant « tout sauf une pollution ordinaire », un calculateur spécifique a dû être pensé (The Federal Interagency Solutions Group, 2010, p. iii). Ce calculateur a été conçu par deux équipes

de scientifiques[36]. Au total, vingt-sept contributeurs et quatorze relecteurs ont été impliqués dans l'élaboration de ce calculateur. L'enjeu des équipes de scientifiques et d'experts était de déterminer la masse totale des hydrocarbures relâchés par le puits en fonction des formes qu'ils ont prises dans l'environnement. On parle de processus de *weathering*.

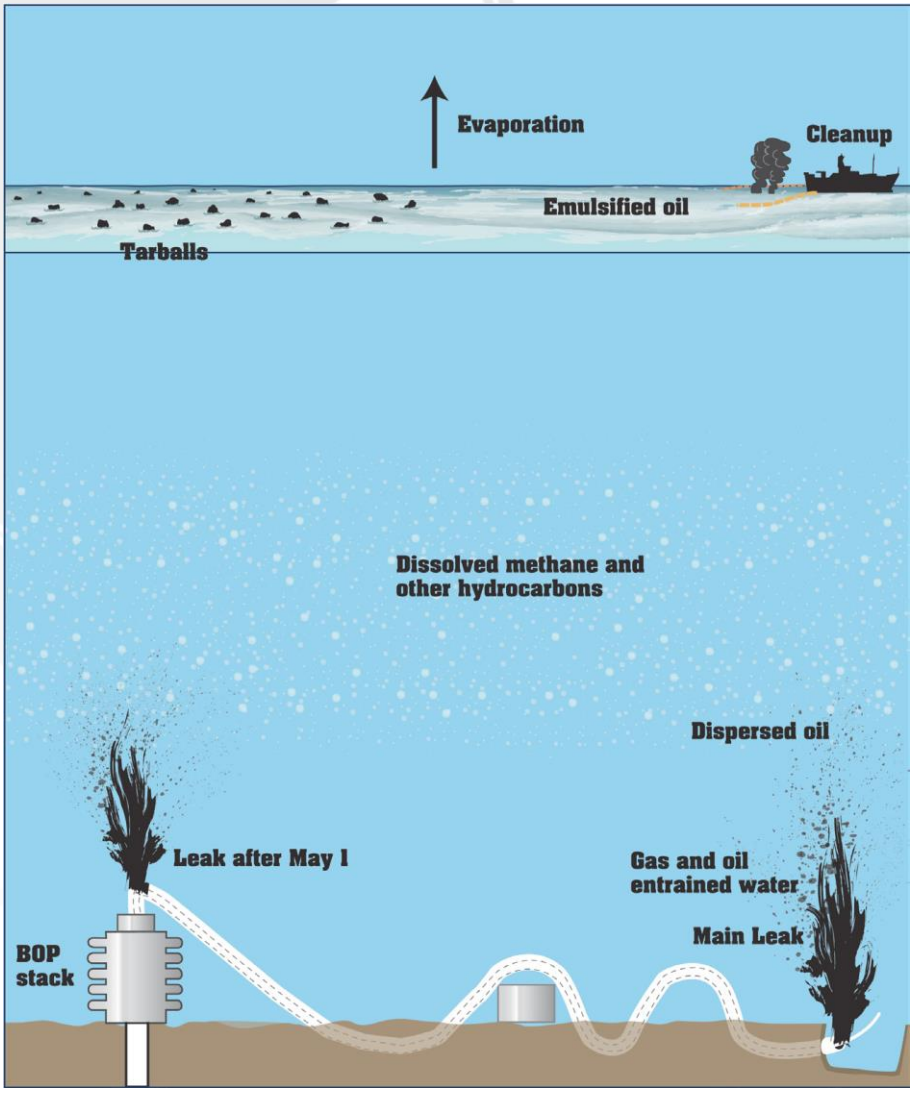


Figure 24 : les différentes formes que prend l'effluent en sortie du puits (source (The Federal Interagency Solutions Group, 2010, p. 3))

A l'aide de formules mathématiques qui modélisent les différents comportements de l'effluent dans l'environnement, les équipes sont capable de proposer une estimation du volume relâché. Les résultats sont en page trente-neuf du rapport ; on a un ensemble de trois valeurs selon trois conditions : *best case*, *expected* et *worst case*. *Best case* et *worst case* sont les combinaisons de valeurs des sept variables représentées dans chaque histogramme empilé et qui correspondent aux valeurs extrêmes (inférieure et supérieure de l'intervalle de confiance à 95 %) pour l'estimation

36 L'équipe *Flow Rate Technical Group/Department of Energy* et l'équipe *United States Geological Survey/National Oceanic and Atmospheric Administration/National Institute of Standards and Technology*.

du volume des *other oil*, lui-même étant le paramètre estimatif de l'incertitude du calculateur.

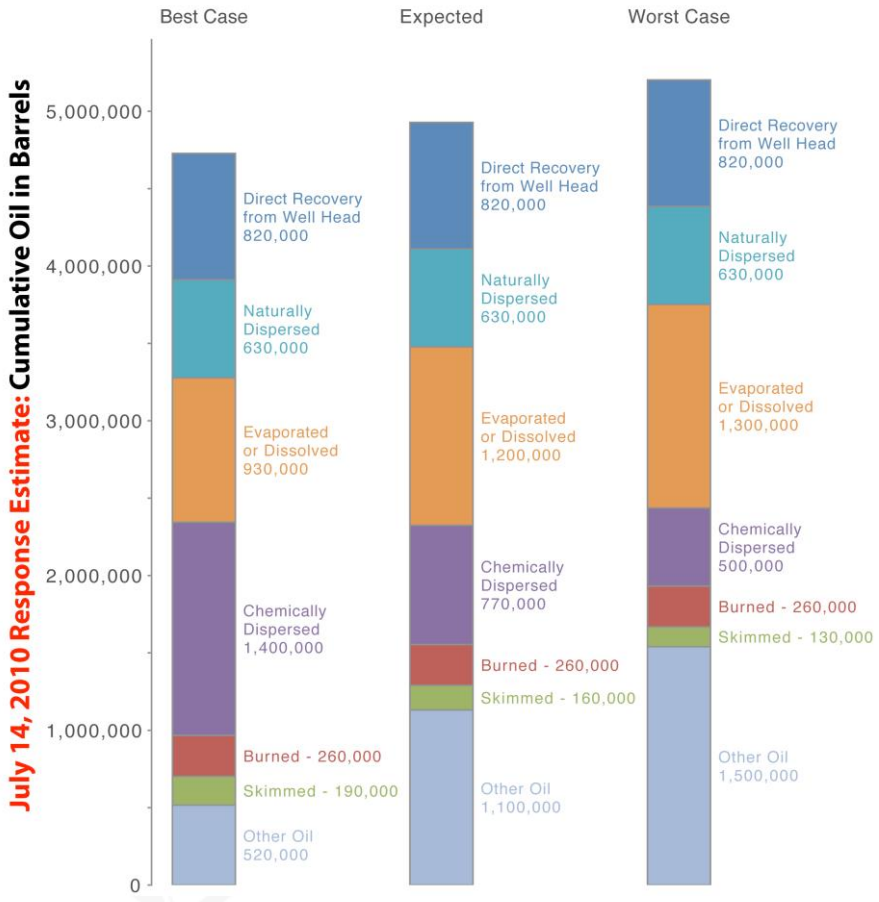


Figure 25 : les trois histogrammes empilés représentant trois valeurs possibles du volume total relâché

On a donc au total :

- valeur *best case* : 4 750 000 bbl ;
- valeur *expected* :4 940 000 bbl ;
- valeur *worst case* : 5 140 000 bbl.

Le calculateur fut rendu opérationnel le 22 juin 2010, soit soixante-trois jours après la survenance du *blowout* au 20 avril.

Comment les experts ont-ils fait avant pour évaluer la fuite ? Nous avons recensé et comparé l'ensemble des valeurs qui ont été proposées dans différentes sources[37] pour l'estimation du débit de fuite, qui *a fortiori* détermine le volume relâché. Nous

37 En particulier les papiers (Camilli *et al.*, 2012; Lubchenco *et al.*, 2012; McNutt, Camilli, *et al.*, 2012; McNutt, Chu, *et al.*, 2012) et les rapports (Allen, Thad W., 2010; National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, 2011a; United States Coast Guard, 2011).

avons dressé un graphique qui illustre le degré de connaissance, au jour le jour, pendant l'intervention, à propos de ce débit (mesuré en *Barrel Per Day*, *BPD*) :

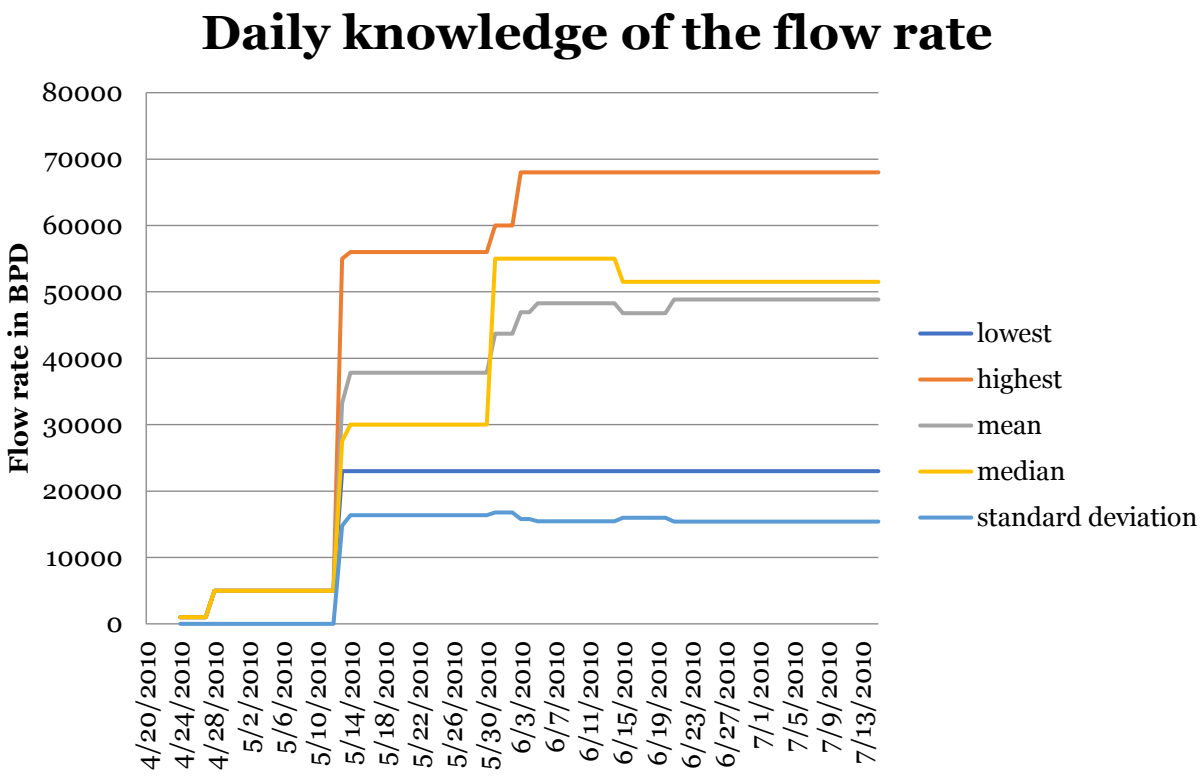


Figure 26 : degré de connaissance quotidien du débit de fuite

Au cours des vingt-trois premiers jours, les intervenants n'ont aucune idée « réelle » du débit. Nous recensons ci-dessous les quelques valeurs qui ont été proposées et dont il n'a pas été possible de trouver une fondation scientifique pour les étayer :

- 24 avril, 1000 *BPD*, donnée par BP ;
- 25 avril, un doute est émis sur 1000 *BPD* par les intervenants ;
- du 26 au 28 avril, entre 1000 et 6000 *BPD* ont été annoncés par BP ;
- le 26 avril, la valeur de 5000 *BPD* est estimée par un scientifique de la *NOAA* extérieur à l'intervention qui a informé un autre scientifique de la *NOAA* intégré au dispositif ;
- le 28 avril, l'amiral Landry (alors *FOSC*) a déclaré : « *les experts de la NOAA pensaient que la production pouvait atteindre 5000 barils [par jour]* ».

Jusqu'aux premières mesures lancées par l'équipe de l'université de Berkeley le 13 mai, l'une des équipes de scientifiques intégrées au sein du *FRTG* (créé à la demande du *NIC)*, le débit officiel restera de 5000 *BPD*. Ensuite, différentes équipes avec différentes méthodes vont proposer des estimations du débit de fuite jusqu'à la fermeture du puits le 15 juillet 2010. Les valeurs moyennes et médianes sont proches et se rapprochent au fil du temps. L'écart-type est régulier, mais il est assez élevé, illustrant une dispersion importante des valeurs proposées par les scientifiques.

On voit donc deux choses : les équipes d'intervention n'ont pas de connaissance objective du débit pendant plus de trois semaines et la détermination de ce débit a été une tâche difficile à accomplir et qui n'a pas abouti non plus à un consensus marqué.

1.3.2.2 Le processus pendant le procès

Dans le document intitulé *Phase deux des constatations de faits et conclusions de droit* (Judge Barbier, 2015, p. 37)[38], on retrouve une série d'expertises qui proposent un ensemble de valeurs des estimations de l'attribut *oil budget*. Quatre experts mandatés par les États-Unis, deux mandatés par BP ont eu à quantifier l'attribut *oil budget*. Le troisième acteur, la Cour fédérale, présidée par le juge Barbier interviendra par la suite. En page trente-neuf du document, on présente les différents experts et leurs estimations : six ensembles de valeurs pour quinze valeurs au total. Cela varie tout de même du simple au double. Les paragraphes 262, 263 et 264 illustrent encore une fois la construction sociale des faits : si l'on se penche encore plus en profondeur, on a au paragraphe 265 et aux suivants l'explication de l'élaboration d'un rapport d'expert et on y retrouve les mêmes processus et la nécessité de faire de ces documents des « forteresses imprenables » avant de les soumettre à la controverse extérieure, en l'occurrence le camp adverse et la Cour de justice. Une fois ces rapports soumis, l'administration et BP proposent chacun à la Cour une valeur à l'attribut. Enfin, la Cour, représentée par le Juge Barbier retiendra deux valeurs, ou plutôt une différence : elle considérera qu'il y a quatre millions de barils sortis du puits, mais que l'effort mené sur la source (le puits) a permis d'en récupérer 810 000 bbl. La valeur de 3 190 0000 bbl effectivement relâchés dans l'environnement sera donc retenue contre BP. L'attribut *oil budget* de l'évènement *Deepwater Horizon* est donc représenté par dix ensembles de valeurs, soit vingt-deux valeurs présentées dans l'histogramme :

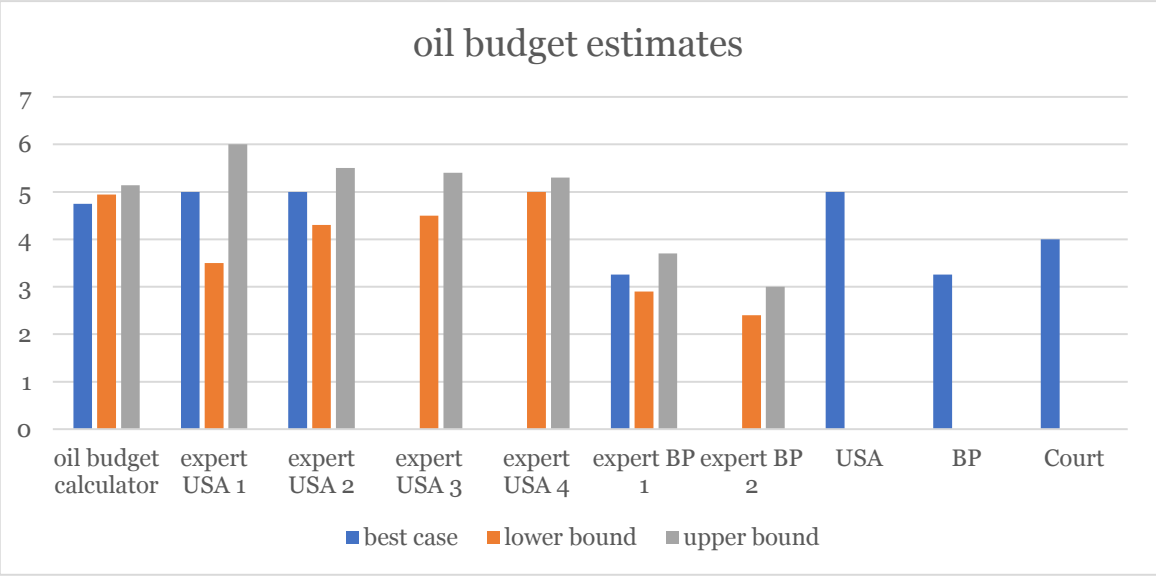


Figure 27 : les représentations de l'attribut *oil budget*

Là aussi, il n'y a aucune raison qu'il y ait consensus. L'estimation d'une valeur telle que la quantité d'hydrocarbure relâchée par le puits pendant l'accident confronte le chercheur à la multiplicité des sources et des résultats. La valeur de quatre millions de barils (sortis du puits) retenue en conclusion du procès semble finalement être la

38 Le procès-verbal de la phase 2 du procès États-Unis contre BP, le plus important contentieux juridique à l'heure actuelle de ce pays.

« traduction » numérique, au sens de Callon (1986), d'une représentation sociale de l'accident élaborée au fil des discussions par BP, les États-Unis et la Cour de justice.

1.3.3 De la nécessité d'organiser les données aux fins d'une connaissance scientifique

L'exemple précédent interroge quant à la valeur à attribuer aux informations disponibles pour une démarche de recherche. La littérature sur la qualité de la donnée dans le champ des *safety studies* est assez maigre. Qureshi (2008, para. 6.2), tout en proposant un examen très approfondi des approches de modélisation des accidents, écrit seulement un paragraphe qui traite de l'importance de la qualité du rapport d'enquête. Tout au plus, Psarros et al. (2010) soulignent le danger d'utiliser des données historiques (sur les accidents maritimes) sans remettre en question la fiabilité de ces mêmes bases de données. La quantité de données nécessaires et les choix afférents quant à leur traitement ne sont pas abordés non plus. Pourtant, la justification du choix des données est aussi importante pour la validité des résultats que la qualité du raisonnement qui s'en suit. Nous posons donc la question suivante : où sont les justifications des auteurs quant à leurs choix de données générées après un accident et leur positionnement méthodologique ? Il semble qu'il y ait un manque flagrant de réflexion sur la qualité et la quantité des données utilisées dans le domaine des *safety studies*. Nous proposons ci-dessous une représentation du flux de données entre l'objet « accident » et la science en passant par l'investigation technique officielle :

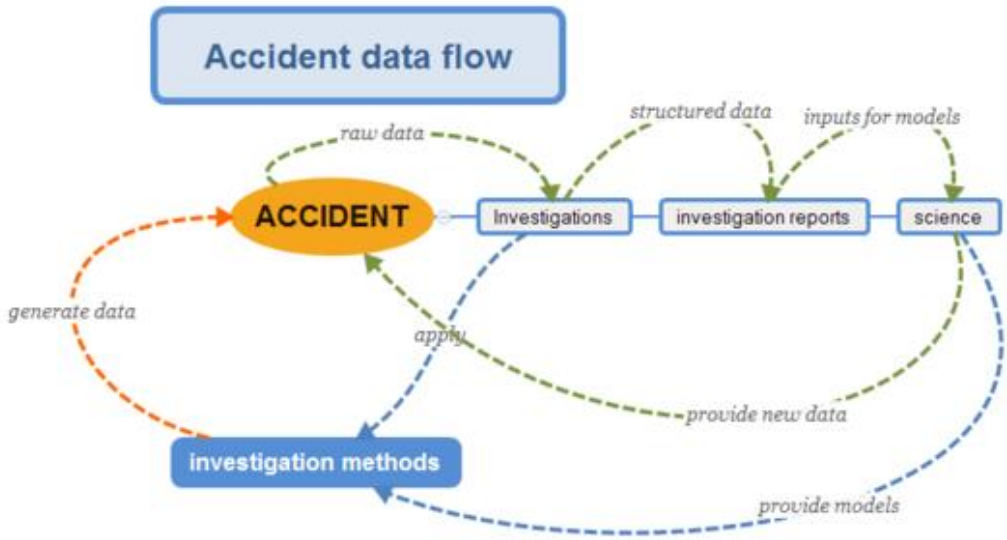


Figure 28 : le flux des données liées à un accident

Cependant, il existe des domaines scientifiques où la qualité des données semble scrupuleusement surveillée et contrôlée. Dans le domaine médical, et plus particulièrement dans les soins apportés au patient et son ressenti, des travaux ont interrogé la fiabilité et la validité de la multitude et de la diversité des données et en particulier des *medical charts reviews*, un document intermédiaire entre le patient et

le soignant. Ils ont montré de nombreuses lacunes méthodologiques dans leur élaboration et dans les informations contenues. Ces documents étant à la base de très nombreux travaux de recherche, cela amène *in fine*, à douter des conclusions tirées de ces mêmes travaux comme le rappellent Eder et al. (2005), Gilbert et al. (1996) et Gregory and Radovinsky (2012). La question se pose également quant aux données médicales directement obtenues via le patient comme les *Patient Reported Outcomes (PRO)* (Frost *et al.*, 2007) ou le niveau de satisfaction des soins (Sitzia, 1999). Il semble que les différentes voies d'accès à la donnée dans le domaine médical soient particulièrement surveillées et évaluées, le problème est cerné. Une recherche remarquable d'exploitation des *PRO* dans une démarche de modélisation du patient face à la douleur chronique est menée aujourd'hui par l'université de Stanford[39]. Elle montre l'intérêt pour la science de se préoccuper de la qualité des données utilisées dans ce domaine.

Les données issues d'un accident sont très nombreuses, diverses et de qualité variable. Autant de nombreux auteurs ont mis en garde quant à la qualité de l'élaboration des rapports d'enquêtes, *via* notamment la mise en œuvre des méthodes d'enquêtes, autant la qualité de ces données n'est pourtant que rarement interrogée dans le champ des *safety studies,* surtout lorsque le rapport d'enquête devient matériau de recherche. On voit également que pour un même phénomène, de nombreuses représentations peuvent être proposées sans obligation quelconque de « convergence ». A cela s'ajoute la très grande quantité de sources fournissant des données suite à un accident, pour certaines très au-delà des cercles institutionnels et académiques.

Nous proposons pour conclure ce chapitre de préciser les concepts de données, d'information et de savoir, que nous envisageons d'organiser par la suite de manière systématique. Nous partons du modèle *Data Information Knowledge Wisdom*, ou modèle *DIKW* :

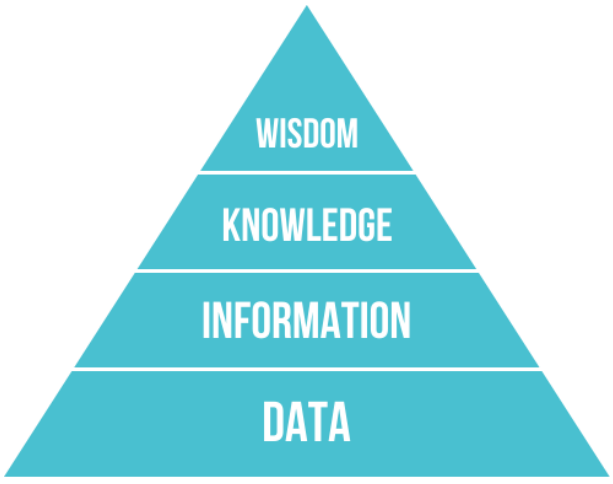


Figure 29 : le modèle DIKW

39 Source : https://choir.stanford.edu/clinical-practice/

Il semble que les premiers théoriciens à proposer ce modèle soient Zeleny (1987) et Ackoff (1989)[40]. Il a été à de très nombreuses reprises revu, commenté, critiqué et enrichi[41] (2007).

La hiérarchie de cette pyramide fait consensus (Rowley, 2007b, p. 174).

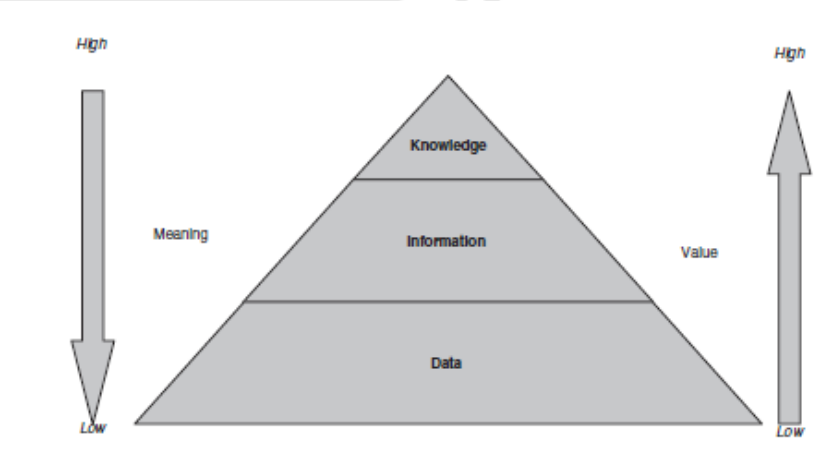


Figure 30 : extrait de (Chaffey et Wood (2005), cités dans Rowley, 2007b, p. 5)

Plus on monte dans la pyramide, plus il y a de la valeur en corrélation avec la signification apportée (Ackoff, 1989). Par construction, les données servent à créer de l'information qui sert à créer de la connaissance qui permet d'accéder à la sagesse (Rowley, 2007). Pour ces auteurs, les données sont des ressources, des objets manipulables (Hey, 2004, p. 6), que l'on peut empiler et quantifier. Les données sont discrètes, des faits objectivés ou des observations qui ne sont ni organisées ni traitées ; elles ne portent aucune signification particulière (2007, p. 170). Rowley (2007, pp. 174–175) soulève le fait qu'il n'y a pas de consensus autour du fait de savoir si c'est la structure qui amène du sens aux données ou si c'est le récipiendaire qui, lorsqu'il accède à une donnée, va amener sa propre signification pour en faire une information. Nous voyons les données comme une source. Cet auteur considère en outre que l'information est le résultat d'un processus qui confère aux données « *une fin ou un contexte précis, ce qui les rend significatives, précieuses, utiles et pertinentes* » (*Ibid.*). L'information crée le lien entre les données.

La connaissance est le pouvoir de répondre à la question « comment ? » (Ackoff, 1989). Elle n'existe pas *a priori*, mais elle permet, en étudiant sa construction, de dégager des « invariants méthodologiques » (Le Moigne, 1995 - cf. 5.1, cité dans Charlet, 2002, p. 19). La connaissance est créée sur des considérations sémantiques fondées sur les hypothèses sur le monde que nous nous faisons (Jashapara et Newell et al, cités dans Rowley, 2007, p. 173).

La compréhension, c'est apporter la réponse à la question « pourquoi ? », c'est donner une explication (Ackoff, 1989). Il est intéressant de noter que plus on s'élève dans la hiérarchie de la pyramide, plus les positions divergent au sujet des concepts qu'elles évoquent : la compréhension, pour Bellinger et *al.* n'est pas un « étage de

40 Il semble qu'une une proposition antérieure aurait été présentée dans une parution de la *World Future Society*, https://wfs.org/, mais nous n'en avons pas trouvé la trace.
41 Voir les travaux notamment de Jifa et Lingling (2014), Bernstein (2011), et Nürnberger et Wenzel (2011).

plus », mais bien un processus qui intervient à tous les étages pour passer à un étage supérieur (2004). Pour Ackoff, la compréhension permet à l'humain d'améliorer son efficience. C'est le rapport entre la mobilisation (de ressources, d'énergie, de temps) et la possibilité d'atteindre l'objectif fixé. L'efficacité quant à elle, se mesure au prisme de la valeur de l'objectif (1989).

Enfin, pour conclure sur cette présentation du modèle *DIKW*, Rowley souligne qu'il y a peu de littérature dans le domaine des systèmes d'information, du management de la connaissance et plus généralement des sciences de l'information qui prend en considération le haut de la pyramide : la sagesse. Pour Ackoff, la sagesse est liée aux valeurs, elle exige la capacité de jugement. Par opposition aux étages inférieurs qui peuvent être informatisés, la sagesse est propre à l'espèce humaine. Concernant nos travaux sur l'accident de *Deepwater Horizon*, le champ lexical de la sagesse (*wisdom*) a été employé par un géologue de l'USGS, le Dr. Paul Hsieh, pour qualifier le *leadership* du chef du *GLST (Government-Led Science Team)*[42], le Pr. Steven Chu. Nous présentons notre traduction d'un extrait d'un entretien que nous avons mené en avril 2016 avec le Dr. Hsieh ; il est question de la prise de décision :

> « *[…] parce que, comme je l'ai dit, si vous suivez les considérations rationnelles exclusivement, comme je l'ai dit, si vous ne voulez pas être tué dans un accident de voiture, ne conduisez pas. Mais prendre une décision demande d'aller au-delà. Et la façon dont vous prenez cette décision est fondée sur la sagesse. Je ne sais pas quel autre mot… Il y a… Vous avez besoin d'un leader sage et pas seulement de quelqu'un qui est absolument, techniquement compétent, expert parce que vous avez besoin de plus que cela au niveau supérieur. Il faut beaucoup de ces compétences au niveau du personnel, mais le meilleur (« the top person ») doit avoir de la sagesse.* »

Ce témoignage illustre, face à des décisions en situation extrême, que la connaissance ne se suffit pas et que la sagesse n'est pas considérée ici comme quelque chose de foncièrement rationnelle.

Nous venons de présenter l'univers dans lequel nous nous inscrivons au travers du prisme de notre matériau de recherche, c'est-à-dire les rapports d'enquêtes et plus généralement les documents créés lorsqu'un accident survient. Il nous fallait situer notre matériau de recherche dans un cadre théorique existant pour nous permettre maintenant de montrer ce que nous allons en faire.

42 L'équipe de scientifiques de très haut niveau chargée de réfléchir aux solutions à mettre en œuvre pour reprendre le contrôle sur le puits.

Nous parlerons à partir de maintenant du processus *DIKU(W)* comme de la représentation de notre épistémologie de la connaissance et nous le représentons de la sorte :

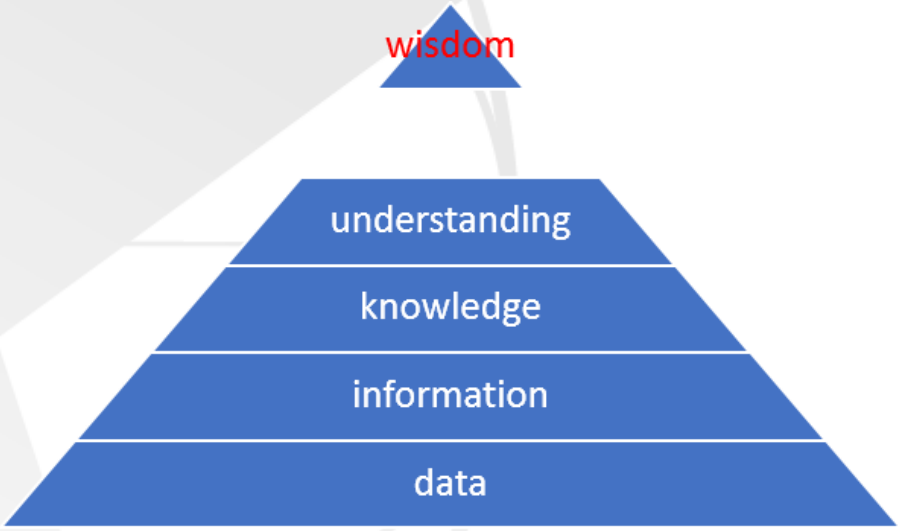


Figure 31 : le processus DIKU(W)

Nous nous contenterons d'atteindre la compréhension, laissons la sagesse à d'autres. Nous allons présenter maintenant notre approche pour la recherche de la compréhension du cas *Deepwater Horizon*.

*

* *

Dans ce chapitre, nous avons montré que le cas *Deepwater Horizon* n'a été abordé par la science que sous quelques angles particuliers (détermination des causes du naufrage d'un côté, évaluation des effets de la « marée noire » de l'autre). La gestion de la crise provoquée par la « marée noire » a été occultée. Cette « situation extrême » (Travadel and Guarnieri, 2015), qui s'est prolongée du 23 avril au 15 juillet 2010, correspond pourtant au « domaine d'existence » dont la frontière marque la transition accidentelle à partir d'une opération normale (cf. notre définition de l'« accident », et la figure 12, section 1.1). L'« accident » de *Deepwater Horizon* ne peut se définir sans référence à cette situation. A ce stade, la « connaissance » sur l'accident est donc parcellaire. Plus largement, il suit des remarques sur le cas faites au chapitre précédent un triple constat.

Premièrement, la qualité des données, leur quantité, leurs supports et leurs traitements sont divers et hétérogènes, et l'élaboration et l'utilisation de la connaissance issue de ces données sont discutables.

Deuxièmement, notre définition du concept d'accident, et la caractérisation du cas *Deepwater Horizon*, invitent à ouvrir le champ des connaissances associées, pour mieux caractériser ses diverses composantes.

Troisièmement, les modèles d'accident disponibles, centrés sur des causes de dommage ou des conséquences, ont vraisemblablement contribué à restreindre la portée des analyses du cas.

Il nous faut donc aborder à la fois le recensement de données, leur mise en relation conceptuelle pour faciliter l'interprétation des différents aspects d'un accident et en proposer une représentation adéquate à une utilisation pour des fins scientifiques.

Chapitre 2 Connaissances, ingénierie, ontologies

Il semble que s'aventurer dans le domaine de l'ingénierie des connaissances soit une piste intéressante pour la résolution de nos problèmes. En effet, l'ingénierie des connaissances s'est attaquée de front à la question des connaissances en usage, aux problématiques de leur partage et de leur diffusion (Teulier, Charlet and Tchounikine, 2005, p. 14). L'ingénierie des connaissances est une science récente[43], apparue à la fin des années 1990. Charlet, (2002, p. 17), en propose la définition suivante :

> « *L'Ingénierie des connaissances se situe au carrefour de plusieurs réflexions : la linguistique pour étudier la formulation linguistique des connaissances ; la terminologie et les recherches sur la genèse des ontologies pour dégager les concepts ; la psychologie pour élaborer les méthodes d'élicitation et les modèles d'assistance ; la logique pour élaborer les modèles formels ; l'informatique pour les opérationnaliser ; l'ergonomie pour interpréter et s'approprier le comportement du système ; les sciences de gestion pour concevoir et replacer les systèmes dans leur environnement organisationnel, etc. Ces réflexions, développées au sein de disciplines, se fécondent réciproquement et conduisent à faire évoluer le paradigme.* »

Teulier et al. (2005, p. 14) considèrent l'ingénierie des connaissances comme le « génie des procédés » où le produit serait la connaissance avec toutes les capacités de traitement de cette matière, notamment la modélisation et tous les « dispositifs professionnels » permettant de produire des résultats. Quoi qu'il en soit, ce sera la discipline dans laquelle nous allons évoluer pour la résolution de notre triple problème.

Le chapitre 2 est la présentation des théories que nous allons convoquer pour faire face aux défis soulevés dans le chapitre précédent. Nous allons en premier présenter le concept d'ontologie (2.1). Nous nous intéressons d'abord à son origine, philosophique, et à son appropriation par la science et en particulier dans le domaine de l'ingénierie des connaissances (2.1.1). Nous nous intéressons ensuite aux critères de conception d'une ontologie scientifique (2.1.2) et nous exposerons les intérêts de ces représentations de la connaissance (2.1.3). Une fois le concept d'ontologie présentée, nous nous pencherons sur leur élaboration et leur production (2.2). Les fondamentaux de logique seront abordés (2.2.1), puis nous introduirons les opérations possibles et leurs intérêts en matière de représentation des connaissances (2.2.2). De là, nous présenterons l'implémentation informatique actuelle, notamment le langage *OWL*, qui permet la réalisation et l'utilisation effectives des ontologies (2.2.3). Enfin, nous allons présenter les critères de choix d'une ontologie pour l'étude du cas *Deepwater Horizon* (2.3). Nous passerons d'abord sur quelques réalisations de qualité en ingénierie des connaissances qui exploitent des ontologies à des fins d'illustration (2.3.1). Nous

43 Voir à ce sujet l'article fondateur de Studer et al. (1998).

présenterons ensuite comment nous avons procéder pour choisir l'ontologie que nous allons utiliser par la suite pour résoudre notre triple problème exposé dans le chapitre 1 (2.3.2). Enfin, nous présenterons en détail l'ontologie de haut niveau *DOLCE,* dont l'utilisation nous permettra de proposer notre ontologie de l'accident (2.3.3).

2.1 Le concept d'ontologie

L'ontologie, qui est le sujet principal de ce chapitre, articule les problématiques que nous avons soulevées à propos des données, de la caractérisation des objets d'étude et des moyens de représenter les connaissances à leur sujet. Nous partirons de l'origine philosophique de ce concept et nous verrons ce que la science en a fait.

2.1.1 D'une ontologie à l'autre

Nous abordons ici succinctement le passage d'une conception philosophique de l'« ontologie » à une conception scientifique. Cette question a d'autant plus d'importance que beaucoup d'auteurs de la littérature dans le domaine de l'ingénierie des connaissances et plus généralement en science de l'information assument peu ou pas la provenance philosophique du concept.

2.1.1.1 L'ontologie au sens premier

Le CNRTL définit l'ontologie comme suit : « Partie de la philosophie qui a pour objet l'étude des propriétés les plus générales de l'être, telles que l'existence, la possibilité, la durée, le devenir »[44]. « Son étymologie vient du grec « to on » à partir du participe présent substantivé du verbe être « einai » et des mots « logia » (théorie) et « logos » (discours), l'ontologie est la « science de l'être en tant qu'être » » (Roche, 2005, chap. 4.2).

La référence première à l'ontologie est philosophique et particulièrement aristotélicienne. C'est l'étude de l'être dans ce qu'il est et dans son évolution avec une recherche de généralités à son propos. Depuis la métaphysique aristotélicienne, l'ontologie traverse les époques et, même confrontée à l'idée de Dieu, les théologiens sont obligés de reconnaître que Dieu aussi a une existence propre et qu'on ne peut donc passer à côté de la compréhension de « la raison intrinsèque de l'existence de l'être » (2007). Deux approches de l'ontologie se dégagent. D'une part, la recherche de la raison intrinsèque qui permet de donner un nom à des objets existants par le biais de l'observation et de la réflexion sur le monde ; d'autre part, la recherche de l'essence, vue par le prisme de la classification et de la hiérarchie des concepts et des propriétés qui permettent la description du monde dans le but de trouver ce qui fait cette essence. La deuxième approche influencera considérablement les champs disciplinaires qui ont fait appel à l'ontologie notamment pour la représentation des connaissances. A ce propos, Robert (2006, p. 59) précise qu'il faut rechercher par essence d'un objet les caractères « eidétiques » comme étant « les primitives participant à la définition d'un concept ».

Dans le dictionnaire de philosophie de l'université de Stanford (Hofweber, 2014), l'ontologie est proposée comme une série de propositions relatives en premier lieu à affirmer ce qui est et proposer une constitution de ce qui est dans une réalité ; puis à chercher à décrire les caractéristiques et relations les plus générales qui existent dans ce que l'on considère comme être. Sauf qu'apporter des réponses à ces

44 (Définition A. – PHILOSOPHIE 1. [L'ontologie] a) [Au xviies. et p. réf. à la philos. aristotélicienne]).

questionnements pose un premier problème qui concerne l'engagement ontologique et un second qui est du ressort de la définition même de ces problèmes, la méta-ontologie.

Hofweber propose une vision de l'ontologie que nous qualifierons d'approche des quatre piliers :

1. « (O_1) l'étude de l'engagement ontologique, c'est-à-dire ce à quoi nous ou les autres sommes engagés ;
2. (O_2) l'étude de ce qui est ;
3. (O_3) l'étude des caractéristiques les plus générales de ce qui est, et comment ces choses sont en relation les unes avec les autres de la manière métaphysique la plus générale ;
4. (O_4) l'étude de la méta-ontologie, c'est-à-dire dire quelle tâche la discipline de l'ontologie doit viser à accomplir, le cas échéant, comment les questions auxquelles elle vise à répondre doivent être comprises et avec quelle méthodologie on peut y répondre. » (Notre traduction, Hofweber, 2014, chap. 3.1. Different conceptions of ontology)

Ces quatre piliers, les « quatre O », structurent l'ontologie comme une discipline : O_4 est le « méta-descripteur » : il explicite la manière dont les trois autres doivent être compris. O_1 est le « rassembleur », la ligne de conduite, celle qui nous pousse à converger autour d'une question qui trouverait son sens dans O_2, qui est le « questionneur », orienté donc par O_1. Si jamais la question proposée par O_2 n'est pas satisfaisante, alors il faut revoir comment O_1 a été établi. Enfin O_3 est le « descripteur » : il ne tient que si O_2 est établi. O_3 sans O_2 (et *a fortiori* O_1) ne serait que spéculatif.

Ces quatre piliers constituent notre référentiel de compréhension de la nature, de la structure et de l'existence des ontologies ; nous articulerons notre réflexion par la suite autour de ces piliers. Nous allons voir maintenant comment cette partie de la philosophie est arrivée dans la science et comment elle a été appréhendée.

2.1.1.2 Ingénierie des connaissances et ontologie

Poli (1999, p. 20) résume toute la difficulté de la coopération entre disciplines et est très critique sur la position des philosophes actuels à propos de leur rapport à la science où il considère qu'ils se sont égarés dans une impasse.

Le court article de Bénel (2011) nous apprend que l'apparition de « l'ontologie », hors considération philosophique, date de la disparation du cercle de Vienne (Chapuis-Schmitz, 2004) en 1936 et que, Wüster, chef d'entreprise de son état, cherche à supprimer les problèmes de communication entre ingénieurs de langues différentes. En 1979 est publiée à titre posthume « La théorie générale de Terminologie » qui y est présentée comme une « ontologie ». Bénel fait remarquer qu'il semble que la communauté scientifique, et particulièrement celle du Web sémantique (nous allons y venir) rechigne à assumer la paternité en philosophie. Pourtant, si l'on se penche maintenant sur les travaux menés dans le domaine de l'ingénierie des connaissances, plutôt dans le domaine de l'intelligence artificielle et des systèmes experts, on trouve un papier de McCarthy (1980, p. 31), l'un des fondateurs du concept d'intelligence

artificielle, qui utilise une fois le terme « *ontology* » en précisant entre parenthèses sa signification comme « les choses qui existent » Ici, la définition est très succincte, mais il y a un rapport évident avec la définition philosophique. Il y a également l'article intitulé « *Enabling technology for knowledge sharing* », dans lequel un groupe de chercheurs (dont Gruber) utilisent le terme « *ontology* » et le définissent comme « *les termes de base et les relations qui composent le vocabulaire d'un domaine d'intérêt particulier ainsi que les règles pour combiner les termes et les relations afin de définir les extensions de ce vocabulaire.* » (Notre traduction Neches *et al.*, 1991, p. 40).

Dans ce dernier cas, la définition est centrée sur l'utilisation qui en sera faite et sur une potentielle structure (notamment la notion d'extension que nous allons aborder par la suite). Deux ans après, le papier séminal de Gruber (1993a) intitulé « *A translation approach to portable ontology specifications* » propose ce qui sera considéré comme fondateur par la communauté scientifique pour la conception d'ontologies[45]. Selon Genesereth and Nilsson (1987) que l'auteur cite, à la base de toute connaissance, il y a un processus de conceptualisation décrit comme une vue simplifiée, abstraite du monde que l'on souhaite représenter dans un but précis. En rappelant que le terme est emprunté à la philosophie, Gruber (1993a, p. 1) affirme qu'une : « *ontologie est un recensement systématique de l'existence* », et amène l'ontologie comme « *une spécification explicite d'une conceptualisation.* »

Cette définition, très partagée, mais aussi critiquée et enrichie, fera entrer l'ontologie dans la communauté scientifique.

Poli (1999) fait le point sur la situation des ontologies dans la communauté scientifique et propose un état de l'art et des questionnements sur l'utilité, les définitions et les controverses liées à l'ontologie. Il propose non pas une définition commune ou unique, mais un faisceau de définitions pour l'ontologie et la sienne comme théorie des objets quel que soit leur « type ». Par la suite, Borst, dans sa thèse, propose l'ontologie comme une conceptualisation partagée (1997, p. 12). Il est intéressant de noter que Borst amène très tôt l'idée de partage, d'échange. Ce que reprend plus récemment Smith qui compare l'ontologie à un réseau téléphonique où l'on retrouve l'échange d'information, mais aussi et surtout l'idée d'accord, de consensus entre les utilisateurs dans la structure et l'utilisation du réseau (2006, p. 1).

Cependant, Noy et McGuinness (2000, p. 3) constatent qu'il y a de nombreuses et contradictoires définitions des ontologies dans la littérature d'intelligence artificielle et elles proposent l'ontologie comme « *une description formelle explicite des concepts dans un domaine du discours [...], des propriétés de chaque concept décrivant des caractéristiques et attributs du concept et des restrictions sur les attributs [...]* ». Une définition synthétique de l'ontologie telle qu'elle s'inscrit dans le champ de l'ingénierie des connaissances est celle de Roche et al., soit « *une représentation d'une modélisation d'un domaine partagée par une communauté d'acteurs* » (2005, p. 57).

Enfin, sur son site personnel[46], Gruber propose en 2009 dans le champ des sciences de l'information qu'une « *ontologie définisse un ensemble de primitives de*

45 Ce papier est cité plus de 17 000 fois sur Google Scholar en juillet 2017.
46 Source : (*Ontology (Computer Science) - definition in Encyclopedia of Database Systems*, no date).

représentation avec lesquelles il est possible de modéliser un domaine de connaissances ou un sujet de pensée. »

Nous adopterons cette définition « d'un point de vue pratique » pour notre compréhension de l'ontologie et nous préciserons notre pensée à propos de l'engagement ontologique et de son positionnement par la suite. En effet, et nous allons le voir par la suite, nous voulons garder, autant que faire se peut, l'universalité de l'approche ontologique et en ce sens, il n'y pas de domaine, ou plutôt, l'ontologie est indépendante du domaine. Comme l'écrit très bien Smith (2006), si l'ontologie est analogue à un réseau téléphonique, alors l'information qui y circule (ce que se disent les gens) ne conditionne en rien le réseau téléphonique. On le verra, la portée de l'ontologie que nous avons choisie pour résoudre notre problème nous permet tout de même d'espérer qu'elle soit quelque peu indépendante du domaine que nous traitons.

A la lecture de ces différentes définitions, nous ressentons qu'il y a bien une forme de convergence pour une définition commune voire unique du concept d'ontologie, mais on constate pourtant que cela prend plutôt la forme d'un faisceau de définitions similaires et que le consensus n'est pas forcément si prégnant que cela. C'est d'autant plus paradoxal puisque la volonté première de l'ontologie est l'unicité dans la description du monde. La difficulté et les limites semblent venir, comme nous allons le voir, moins de la capacité à concevoir et construire les ontologies qu'à penser le monde que nous cherchons à représenter.

2.1.2 Les critères de conception des ontologies

Dans cette partie, nous allons présenter les critères de conception pour les ontologies. Notre travail est basé sur les propositions de Gruber (1995, 1993b, 1993a ; "Tom Gruber | AI product designer," n.d.) et Guarino (1995, 1997, 1998; 2003; 2009; 2009) et plus généralement les travaux issus du laboratoire *Knowledge systems* de l'université de Stanford[47] et de l'école italienne à la fondation du laboratoire *LAO de* Trente[48].

2.1.2.1 Le design selon Gruber et Guarino, ou comment réussir une ontologie

L'ontologie est un produit d'ingénierie, une réalisation qui passe par la formalisation, « *qui existe de façon déterminée et, par extension, qui est énoncé de façon déterminée, claire, sans équivoque* »[49].

Pour ce faire, Gruber (1995, pt. 3. Design criteria for ontologies) propose un jeu de critères qui se veut objectif dans le sens où il est tourné vers le résultat de cette formalisation, et non vers des *a priori* de vérité ou de nature :

- la « clarté », qui réfère à la dépendance à l'égard du contexte, a la place prépondérante. Il faut privilégier des définitions fermées qui ne peuvent laisser le choix à l'interprétation. Si en plus elles peuvent être traduites en axiomes logiques, alors l'explicitation n'en sera que meilleure. L'ontologie doit être justement nécessaire et suffisante ;

47 Source : (*Stanford Knowledge Systems, AI Laboratory*, no date).
48 Source : (*Laboratory for Applied Ontology*, no date).
49 Source : (Définition A.– *FORMEL : Définition de FORMEL*, no date).

- la « cohérence » renvoie à la logique d'établissement de l'ontologie et des inférences qui en seront tirées. Aucune contradiction ne doit pouvoir émerger, dès la proposition axiomatique, mais aussi dans les résultats issus des inférences. De plus, on doit se garder de créer d'éventuels doutes par l'utilisation du langage naturel pour la description desdits axiomes ;
- l'« extensibilité » mesure la possibilité « d'étendre » l'ontologie en apportant de nouveaux éléments qui ne doivent pas remettre en question l'existant ;
- le « biais d'encodage minimal » renvoie à la possibilité de traduire le monde à représenter vers la construction ontologique de la manière la plus « transparente »[50] possible. « Le code » ne pas doit dicter la marche à suivre, mais au contraire, il doit pouvoir faire exprimer, et en aucun cas limiter ou « forcer le passage », toute la possibilité de conceptualisation de l'« ontologue », le concepteur de l'ontologie ;
- l'« engagement ontologique minimum » désigne un positionnement de l'« ontologue » par rapport à sa manière de décrire le monde. Il doit être le plus « faible » possible, c'est-à-dire qu'une ontologie « idéale » ne saurait « enfermer » le processus de conceptualisation dans une théorie trop rigide qui pourrait *in fine* arrêter tout simplement ce processus, soit par son impossibilité d'achèvement (pas de possibilité d'instanciation), soit par une trop grande « étanchéité conceptuelle » de la représentation pensée par le créateur de l'ontologie (pas de possibilité de spécification ou de portabilité).

Guarino (1995, p. 5), un des pères de la réflexion ontologique dans le champ des *computer science* reprend l'ontologie de Cocchiarella comme le développement axiomatique, formel et systématique de la logique de toutes les formes et modes d'existence. Il s'appuie sur cette définition pour conclure que l'ontologie est plus dans la description des « formes » des individus qui composent un monde plus que de leur existence « nue ». L'ontologie est vue comme la théorie des distinctions entres les « entités » du monde et les catégories descriptives (« *meta-level* ») pour les modéliser. C'est une proposition méta-ontologique (O_4).

2.1.2.2 OntoClean

Guarino et Welty vont apporter une méthode permettant de valider « *l'adéquation ontologique des relations taxinomiques.* » (2009, p. 1). Il s'agit pour les auteurs, à partir de concepts philosophiques, d'évaluer et de valider les choix taxinomiques de l'ontologie et donc de pouvoir proposer des ontologies « claires »[51]. Ils posent une terminologie directement en lien avec la mathématique et la logique pour lever toute forme d'ambigüité qu'il pourrait y avoir dans la réutilisation de certains termes, notamment en informatique.

Les termes explicités sont propriété, intension, classe, prédicat, extension, instances, le principe de subsomption, l'essence, la rigidité, l'identité et l'unité.

50 En anglais, le terme le mieux adapté serait « *seamlessly* ».
51 C'est comme cela que nous comprenons le sens de clean dans ce papier.

La *propriété*, est la signification (ou *l'intension*, au sens logique) d'une expression ; en logique du premier ordre, cela serait un *prédicat* unaire. Cependant, ils insistent sur le fait que la propriété est indépendante de la syntaxe (contrairement au prédicat). Par exemple, si « être marin » et « *be a seafarer* » désignaient la même propriété exprimée dans deux syntaxes différentes (différenciées par la langue), elle devrait conduire à deux prédicats. C'est effectivement la vision conceptuelle la plus juste à notre sens, mais elle est très puriste aussi. Aussi, à partir du moment où l'on implémente une propriété « dans » un prédicat en utilisant systématiquement et rigoureusement le même « logiciel d'implémentation », nous pouvons considérer que c'est de même nature.

La *classe*, ou *extension*, est l'ensemble des entités qui partagent dans le même monde les mêmes propriétés. Les membres de cette classe sont appelés *instances* de cette classe. De la même manière, les classes ne sauraient idéalement être dépendantes de « l'état du monde ». Cette distinction fondamentale peut être comprise de la sorte[52]. Si l'on décide de définir l'ensemble des nombres pairs : on peut proposer deux définitions :

- l'une intensionnelle, telle que : $P = \{n \in N | \exists p \in N, n = 2p\}$;
- l'autre extensionnelle, telle que : $P = \{0,2,4,6,8,10, \dots\}$.

On comprend que la première définition donne la nature de ce qu'est un nombre pair tandis que la seconde désigne ce qu'est l'ensemble des nombres pairs.

Selon Guarino, le *principe de subsomption* s'exprime tel que :

P subsume Q si et seulement si il n′y a aucun modèle tel que Q ∧ ¬P (équation 7)

Cela correspond, dans une logique propositionnelle à une relation du type :

$$Q \rightarrow P$$

Ou, en termes ensemblistes, quel que soit « l'état des affaires », le monde possible, il n'y a aucune instance de Q qui puisse être autre chose qu'une instance de P au départ. Nous reprenons en outre les quatre contraintes associées. Avec p et q deux propriétés et *p subsume q* :

1. *« Si q est anti-rigide, alors p doit être anti-rigide ;*
2. *Si q porte un critère d'identité, alors p doit porter le même critère ;*
3. *Si q porte un critère d'unité, alors p doit porter le même critère ;*
4. *Si q a un critère anti-unique [par exemple le fait d'être indénombrable], alors, p doit aussi avoir un critère anti-unique » (Notre traduction, Guarino and Welty, 2009, p. 6).*

A partir des définitions proposées, les auteurs posent des contraintes basées sur les hypothèses des travaux de Lowe cités par Guarino et Welty (2009, p. 6) : chaque entité doit instancier une propriété la plus générale possible « portant » un critère de son identité.

52 Nous prenons cet exemple du cours du professeur Tellier ('Introduction au TALN et à l'ingénierie linguistiqu.pdf', no date, p. 63).

L'essence est considérée comme une propriété qui doit être vérifiée comme vraie pour toute entité la réclamant dans tous les mondes possibles. Cette essence peut être « *rigide* » : toute instance d'une propriété « rigide » ne pourrait devenir une autre instance dans un autre monde. La « rigidité » est une « *méta-propriété* » qui doit aider l'ontologue à éclaircir son engament ontologique.

Il y a évidemment des propriétés non-rigides, qui au contraire, dépendent de « l'état des affaires du monde possible. » Guarino et Welty, toujours dans le même article, apportent la différence entre propriétés semi-rigides, indispensables à certaines instances dans un monde possible et des propriétés anti-rigides qui ne sont pas essentielles à toutes les instances.

OntoClean propose de détecter des propriétés qui portent le critère *d'identité*, mais reconnaissant que le concept d'identité est difficile à cerner, ils proposent de procéder dans l'autre sens, c'est-à-dire de déterminer les critères nécessaires pour faire *identité* qu'ils appellent « *essential properties* » de cette manière : si l'on n'est pas capable de déterminer ce qui fait identité dans la propriété Q, on pourra beaucoup plus aisément déterminer ce qui fait non-identité avec la propriété P dès lors que toutes les instances de Q ont pour propriété essentielle P.

Quant à *l'unité*, c'est le critère qui détermine les frontières des propriétés. *OntoClean* propose des critères d'unités, c'est-à-dire des critères qui décrivent la manière dont les propriétés sont « unies » : cela peut être topologique, morphologique ou encore fonctionnel. De la même manière, des « touts » peuvent être aussi formés d'autres « touts » unis de manière différentes ou non. Il n'y a aucune prévalence particulière.

Ces critères servent aujourd'hui de « norme » de conception des ontologies, tant sont nombreux les auteurs qui avancent avoir conçu leur ontologie conformément à la méthode *OntoClean*[53].

Ce n'est en revanche probablement pas le cas de l'ontologie à l'origine de la norme ISO 15926 *Life Cycle Data for Process Plant,* pour la standardisation de données de procédés au long du cycle de vie[54]. Cette ontologie est décrite par Leal (2005), son promoteur, et explicitée par Batres et al. (2007). Il y a également un exemple de son utilisation proposé par Fiorentini et al. (2013). Mais Smith (2006) considère cette ontologie comme le contre-exemple d'une ingénierie réussie, tant elle viole les critères précédents.

2.1.2.3 La classification des ontologies

Poli (1999, p. 18) propose une classification des ontologies en quatre catégories :

- ontologie générale ;
- ontologie « régionale » ;
- ontologie de domaine ;

53 Source : (*Barry Smith*, no date).
54 Sources (*15926.org - Home*, no date; *ISO 15926-2:2003 - Industrial automation systems and integration -- Integration of life-cycle data for process plants including oil and gas production facilities -- Part 2: Data model*, no date).

- ontologie d'applications.

Plus récemment, Roussey et al. (2011) proposent deux types de classification :

- en fonction de l'expressivité du langage (un point de vue très logicien et informaticien) et du formalisme (logique et linguistique) ;
- en fonction de la portée de l'ontologie ou de sa « granularité ».

Cette dernière classification rejoint celle proposée par Poli qui est la plus communément établie. En fait, si on s'en tient à l'approche de Guarino ou de Smith, si une ontologie est dépendante du domaine ou avec une granularité plus fine, elle n'est plus une ontologie. Seules les ontologies fondationnelles, « de haut niveau », indépendantes du domaine et de la question (du problème à résoudre) trouvent grâce à leurs yeux, même si en pratique il ne peut s'agir que d'un idéal (cf. *infra*). Quelle que soit la classification que l'on fait, il faut avant tout éviter d'utiliser ce terme pour décrire son objet de travail s'il n'est pas possible de montrer qu'il respecte les critères de conception établis par Gruber ou la méthode *OntoClean* de Guarino et Welty.

On constate que la conception ontologique se formalise plus facilement que l'accord sur une définition commune de l'ontologie. Voyons maintenant quels sont les intérêts de l'ontologie, le pourquoi du comment si l'on ose dire.

2.1.3 Intérêts des ontologies en ingénierie des connaissances

Poli (1999) expose les différentes utilisations des ontologies où elles font consensus que nous résumons ainsi :

- représentation, acquisition, partage, intégration et réutilisation de la connaissance ;
- méthodes de résolution de problème ;
- spécification logicielle ;
- conception de base de données orientées objet ;
- traitement automatique du langage naturel ;
- modèles d'ingénierie ;
- représentation de la réalité et du sens commun ;
- modélisation de la matière, de l'espace, du temps et de la causalité ;
- modélisation des connaissances médicales, des connaissances juridiques.

L'intérêt des ontologies est assez vaste même si on voit que la représentation et l'utilisation (de diverses manières) des connaissances est l'enjeu principal, à travers la crétion des instances, que nous aborderons en profondeur dans le chapitre 3.

Guarino (1995) fait remarquer le manque d'intérêt des *Computer Science* pour les ontologies qui, pour se développer, nécessiterait une « *perspective hautement interdisciplinaire* ». Poli (1999, p. 21) appelle aussi à un rassemblement des sciences autour du concept d'ontologie et, en renversant le point de vue philosophique entre « *prima et ultima philosophia* », amène la science au centre des préoccupations comme une volonté de rappeler à la communauté de s'y consacrer davantage. En 2009, l'arrivée de la méthode *OntoClean* a probablement aidé les ontologies à (re)devenir un

objet d'étude intéressant pour la science et en 2018, nous pouvons penser, au vu des évolutions technologiques récentes en matière de Web sémantique et d'intelligence artificielle, que les ontologies ont été reconsidérées avec beaucoup plus d'intérêt, probablement lié aux enjeux en matière de traitement automatique du langage naturel qui nous semble être l'un des fers de lance de la recherche en la matière.

2.2 Ingénierie des ontologies

Dans cette partie, nous présentons comment sont construites les ontologies dans le cadre de l'ingénierie des connaissances. Nous adoptons une approche résolument pratique de l'ontologie : si la démarche conceptuelle et les considérations logiques sont essentielles à la compréhension et l'utilisation des ontologies, nous nous inscrivons néanmoins dans une démarche d'utilisateur concernant les codes et les logiciels qui permettent le traitement des ontologies. Nous ne sommes pas développeurs et notre démarche a toujours été guidée selon un principe d'efficience : l'objectif à atteindre avec les ressources les mieux adaptées.

Le logiciel que nous utilisons est le logiciel *Protégé*[55], développé par l'université de l'université de Stanford. Il en existe d'autres notamment *Cytoscape*[56] , mais nous avons effectué la majeure partie de notre travail avec *Protégé*. Nous nous sommes appuyés sur le manuel d'utilisation de ce logiciel rédigé par Horridge et al. (2011). Le grand intérêt de *Protégé* est de rassembler une suite d'outils pour la création, l'évaluation, la population, la génération d'inférences, l'interrogation des ontologies écrites en *RDF* ou en *OWL*. A cela s'ajoutent des *plugins*, des extensions de fonctionnalités, qui permettent par exemple des représentations graphiques ou des interfaces de requêtes. De plus, il est possible d'arranger à sa convenance l'interface homme-machine pour améliorer la productivité.

Nous venons de voir quels sont les critères les plus partagés en matière de conceptualisation d'une ontologie. Il s'agit d'une sorte de cahier des charges. Nous allons voir maintenant quels sont les ressorts conceptuels qui permettent la construction des ontologies. Guarino (2009, p. 2) décrit le travail de « l'ingénieur ontologue »[57] comme l'analyse des entités pertinentes et l'organisation en concepts et relations, pour les représenter, respectivement, par des prédicats unaires et binaires. C'est une vision d'ingénieur, de faiseur d'ontologies qui correspond bien à notre approche d'ingénierie des connaissances. Il précise également les problèmes de terminologie entre les différentes disciplines qui peuvent amener à confusion et les choix qu'il vaut mieux faire. Effectivement, et nous y avons été souvent confrontés, la terminologie n'est pas systématiquement équivalente entre différentes disciplines et nous nous sommes astreints à préciser, autant que faire se peut, le cadre théorique dans lequel nous évoluons au moment de notre discussion ou présentation, puis à nous tenir au langage y afférent. Neches et al. (1991) proposent une structure ensembliste, construite sur des classes et leurs relations. Cette structure est la plus adoptée par la

55 Source : (Musen, 2015).
56 Source : (*What is Cytoscape?*, no date).
57 Que nous appellerons « ontologue » par la suite.

communauté[58]. C'est une taxinomie définie selon des axiomes. Voyons dans un premier temps la formalisation logique sur laquelle elle repose.

2.2.1 Structure logique

La construction d'une ontologie s'appuie d'abord sur les fondements de la logique des propositions et celle des prédicats. Nous renvoyons au cours de Amsili (2006) pour une présentation de la logique du premier ordre adaptée aux ontologies. Citons également les enseignements de l'*IRIT*[59] et les cours de Gribomont[60] et de Roussarie[61]. Nous rappelons ici les bases utiles à la compréhension de la structure de l'ontologie que nous allons présenter et utiliser par la suite.

2.2.1.1 Logique des propositions

Soit le syllogisme suivant, dont les deux premières lignes sont appelées les prémisses, la dernière ligne est la conclusion (Amsili, 2006, p. 3) :

- « S'il pleut, il faut ouvrir les dalots,[62]
- Les dalots ne sont pas ouverts,
- Il ne pleut pas ».

Il repose sur une forme logique du type « contraposée » :

$$Si\ P \rightarrow Q, alors\ \neg Q \rightarrow \neg P \qquad \text{(équation 8)}.$$

La logique propositionnelle manipule des propositions. Une proposition est une expression qui peut être dite vraie ou fausse ; en linguistique cela exclut de très nombreuses formes de phrase (impérative, interrogative…). Il est ainsi plus rapide de considérer une proposition comme une phrase déclarative qui ne contient pas d'énoncé modalisé. Mais Amsili précise que l'équivalence entre ce type de phrase et la proposition n'est, en aucun cas, systématique. Comme le souligne Amsili, « *Il faut noter qu'en logique propositionnelle, les propositions vont rester inanalysées, atomiques – sauf quand elles peuvent se décomposer en d'autres propositions et des connecteurs* » (2006, p. 4).

Pour les connecteurs, Amsili les définit comme « *des opérateurs qui permettent, en reliant deux propositions, de former une nouvelle proposition* », *et précise que seuls* « *les connecteurs caractérisés par leur sensibilité exclusive à la vérité ou fausseté de leurs opérandes* » (2006, p. 5) ont un intérêt. Malheureusement, il ne s'agit pas nécessairement des « connecteurs » linguistiques comme les conjonctions de coordination ou de subordination par exemple. Elles assurent certes une connexion entre deux phrases (voir deux propositions), mais ne sont pas pour autant « vérificatrices de vérité ».

---

58 A tel point que notre revue de littérature ne nous a pas permis de trouver un autre type de structure.
59 Source : https://www.irit.fr/
60 Source : (*Pascal Gribomont*, no date).
61 Source : (*Chez Laurent Roussarie*, no date).
62 Source : http://www.cnrtl.fr/definition/dalot

Soit, par exemple, quatre propositions vraies :

a) Mike Williams est électricien et travaille à l'offshore pétrolier.
b) Mike Williams travaille à l'offshore pétrolier parce qu'il est électricien.
c) Mike Williams travaille à l'offshore pétrolier.
d) Mike Williams est électricien.

Si l'on remplace la proposition c) par une autre proposition vraie, par exemple, « le vent souffle », alors la proposition « Mike Williams est électricien et le vent souffle » est également vraie. Si maintenant, on remplace la proposition c) par une proposition fausse, la proposition a) résultant de l'association de c) et d) serait fausse. La conjonction de coordination « et » est ici véri-fonctionnelle, la valeur de vérité de la proposition a) ne dépend que de la valeur de vérité des arguments. Par contre, en considérant la proposition b), si la proposition c) est modifiée (quelque soit sa valeur de vérité), alors la valeur de vérité de la proposition b) est susceptible de changer aussi. « Parce que » n'est en effet pas un connecteur véri-fonctionnel. Nous nous restreignons aux connecteurs véri-fonctionnels. En cela, on profite de la « compositionnalité » : on pourra calculer la vérité d'une proposition en calculant la vérité des propositions qui la composent.

Ce questionnement rejoint complètement la notion d'*entailment* du langage *OWL,* que nous verrons par la suite, et montre bien qu'au niveau du langage, la vérification de la logique n'est pas systématique ; le langage naturel est ambigu[63] et donc il faut trouver une solution à ce problème.

Un arbre de construction permet de vérifier que la formule principale est bien formée dès lors que la décomposition amène à des sous-formules qui sont des formules élémentaires définies dans le langage au préalable. A partir de cette syntaxe, il est possible d'associer une signification à une formule bien formée, dans ce cas, on parle de valeurs de vérité, « vrai » ou « faux ». On associe des valeurs numériques à la valeur de vérité (par convention 1 pour VRAI, 0 pour FAUX) puis on dresse une table de vérité où on pose les valeurs des propositions puis des propositions articulées autour de leurs connecteurs. A ce propos, Amsili rappelle de nouveau que « *la question de la correspondance entre les connecteurs logiques et leurs correspondants directs en langue est, c'est bien connu, délicate* » (2006, p. 8).

Les formules remarquables sont les tautologies dont la valeur est vraie dans toutes les situations et les contradictions où la valeur est fausse pour toutes les situations (on parle aussi de « processus d'absurdisation » pour cette dernière). Les autres formules sont dites contingentes.

2.2.1.2 Logique des prédicats

Le problème majeur avec la logique des propositions est qu'elle n'est pas capable de traiter « l'intériorité » d'un syllogisme et plus généralement d'une proposition.

63 C'est aussi un des postulats de départ de notre étude de l'expression de la causalité dans le chapitre 3.

Prenons le cas suivant :

- si *toutes les plateformes flottent*[64] *(P1),*
- et *les semisubmersibles sont des plateformes (P2),*
- alors : *les semisubmersibles flottent (Conclusion),*

Le syllogisme est valide.

Maintenant :

- si *toutes les plateformes flottent (P1),*
- et *un avion n'est pas une plateforme (P'2),*
- alors : …

Avec ces deux propositions, il n'est pas possible de conclure le syllogisme puisque les deux prémisses sont de « nature » différente. En effet, le fait qu'un avion ne soit pas une plateforme ne le prive pas pour autant de la propriété de flotter et il est impossible ici de le vérifier (on le pourrait si, suivant le critère d'unité proposé par Guarino, on considèrerait que la propriété « flotter » est critère unique de propriété à la classe plateforme). Le raisonnement par syllogisme dans ce cas ne permet pas de valider une conclusion à partir des deux propositions. C'est l'objet de la logique des prédicats.

Continuons avec un exemple :

a) Deepwater Horizon est plus récent que l'Exxon Valdez.
b) L'Exxon Valdez est plus récent que l'Amoco Cadiz.
c) L'Amoco Cadiz est plus récent que le Torrey Canyon.

Dans ce cas, on a trois propositions : *a priori* on peut identifier pour chaque phrase un sujet s et un prédicat P. Le sujet s correspond à une variable x qui peut prendre différentes valeurs en fonction de la proposition. Il faudra donc poser au préalable le prédicat et les variables pour chaque proposition. Si on pose :

$$a) \equiv P(s_{dwh}) \; avec \; s_{dwh} = Deepwater\ Horizon \; et \; P = est\ plus\ récent\ que\ l'Exxon\ Valdez$$

On s'aperçoit que si l'on continue de la même manière avec les autres phrases, on perd le schéma d'inférences qui existe entre les trois propositions. Un prédicat dit *unaire* de la forme $P(s)$ ne serait donc pas suffisant pour ce genre de propositions car on ne peut pas généraliser à l'ensemble des relations « est plus récent que ».

Si maintenant, on pose le prédicat :

$$P' = est\ plus\ récent\ que;\ tel\ que\ s_1\ est\ plus\ récent\ que\ s_2$$

On a la proposition a telle que :

$$a) \equiv P'(s_{dwh}, s_{ev}).$$

64 Ce n'est pas réellement le cas, il y a des plateformes posées sur le sol marin (*jack-up*), mais cela illustre notre exemple.

Avec un prédicat dit *binaire*, on retrouve le schéma d'inférences issu des trois propositions et, en considérant le prédicat P' comme transitif (ce qui est le cas ici), on peut en déduire également que s_{dwh} *est plus récent que* s_{tc} car :

$$s_{dwh} \textit{ est plus récent que } s_{ex} \textit{ est plus récent que } s_{ac} \textit{ est plus récent que } s_{tc}$$

On formulera de la sorte : « *une phrase atomique sera formée avec un symbole de prédicat n-aire (d'arité n) et n constantes* » (2006, p. 16). *C'est l'expression de signification irréductible dans la construction ontologique :*

$P(s, o)$ avec P, prédicat binaire ; s le sujet et o l'objet dans cet ordre (équation 9).

Ce schéma est le fil conducteur de la modélisation logique qui permet l'axiomatisation de l'ontologie.

Passons maintenant au problème de quantification. En effet, comment traiter une proposition telle que : *Tous les hommes sont libres*. On pose le prédicat H unaire pour *être un homme* et on pose le prédicat L unaire pour *être libre* et on réfléchit à leurs relations possibles ; x est la variable en question. Pour amener un éclairage sur l'interprétation de cette phrase, nous utiliserons le carré dit d'Aristote en quantification restreinte à l'univers du discours :

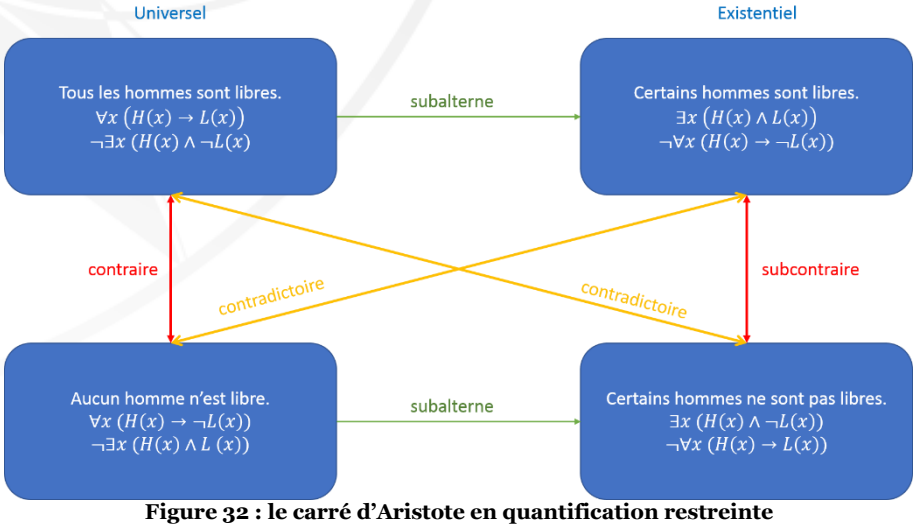


Figure 32 : le carré d'Aristote en quantification restreinte

Pour chaque case, on a exprimé par équivalence en logique des prédicats la phrase en langage naturel. Que veut dire « Tous les hommes » ? Tous les hommes du « monde universel », ou tous les hommes dans un univers « circonscrit », borné ? Pour faire un pont avec les ontologies, Poli (1999, p. 10) propose que l'existentiel soit situé par une contrainte spatio-temporelle tandis que l'universel soit conçu et « reste » dans la pensée. En fait, il existe très souvent un « domaine de quantification implicite » dans lequel s'inscrivent les expressions linguistiques. Il faut donc pouvoir restreindre la quantification à ce domaine en question, l'univers du discours. C'est aussi comme cela que se positionnent les ontologies, comme des ontologies d'universels ou d'existentiels dans la manière donc de considérer l'univers du discours.

Le carré permet de faire apparaître aussi les principes de contraire et de contradiction. L'opération contraire (et subcontraire) est exclusive au domaine, tandis que la contradiction croise les domaines (universel/existentiel). Pour l'opération

contraire, deux propositions contraires peuvent être simultanément fausses, mais ne peuvent pas être simultanément vraies tandis que pour l'opération subcontraire, deux propositions subcontraires peuvent être simultanément vraies, mais ne peuvent pas être simultanément fausses.

Comme pour la logique des propositions, il y a une syntaxe, cette fois-ci articulée autour des formules bien formées et des phrases :

- une formule bien formée est de même nature que dans la logique des propositions ;
- une phrase est une formule bien formée dans laquelle il n'y a pas de variable libre et donc exprime une proposition (au sens rigoureux).

Soulignons que là où la logique des propositions est décidable, c'est-à-dire qu'en un temps fini elle donne un résultat vrai ou faux, celle des prédicats n'est seulement que semi-décidable et une procédure en entrée s'arrête en un temps fini si elle est vraie, sinon elle retourne faux ou ne s'arrête pas. Néanmoins, les deux logiques sont complètes et adéquates ; toute la sémantique est exprimable avec la syntaxe disponible et toutes les formules démontrables sont logiquement valides par rapport à la sémantique du système logique.

2.2.2 Les opérations sur l'ontologie

Une ontologie permet d'effectuer des opérations d'ordre logique, l'inférence et la requête, à partir d'axiomes qui viennent compléter les règles logiques et qui permettent de construire des taxonomies pour ordonner la connaissance.

2.2.2.1 Inférence, requête et axiomes

L'inférence est « *l'opération qui consiste à admettre une proposition en raison de son lien avec une proposition préalable tenue pour vraie* »[65]. Le syllogisme est typiquement une forme d'inférence. L'inférence en soi est un procédé et ne garantit pas la validité de la preuve, son résultat. Il existe des moteurs d'inférences, appelés *reasoners* en anglais, qui déterminent les inférences possibles d'un ensemble de déclarations (propositions) dans une ontologie. Le plus utilisé est le *reasoner HermiT* produit par l'université d'Oxford[66]. C'est celui installé par défaut dans le logiciel de traitement d'ontologies *Protégé*. C'est un domaine en développement et il en existe d'autres comme *Pellet*[67] ou *Fact++*[68] ; et c'est ce dernier que nous utiliserons. Il faut comprendre que derrière le moteur d'inférences, c'est le système de règles logiques qui est en jeu et plus généralement la modélisation logique en elle-même. Le langage de modélisation logique *Description Logics* est une réalisation de ces réflexions menées à ce propos (Baader, Horrocks and Sattler, 2008).

Par ailleurs, les inférences reposent sur « l'hypothèse du monde ouvert » (2011, pp. 69–70), soit le choix de considérer que quelque chose n'existe pas tant que l'on n'a

65 Définition LOG. par le CNRTL (*INFÉRENCE : Définition de INFÉRENCE*, no date).
66 Source : (*HermiT Reasoner: Home*, no date; *HermiT Reasoner: Publications*, no date).
67 Source : (Sirin *et al.*, 2007).
68 Source : ('FaCT++ reasoner | *OWL* research at the University of Manchester', no date; 'List of Reasoners | *OWL* research at the University of Manchester', no date).

pas explicitement déclaré. Autrement dit, si l'on n'est pas capable de déclarer une chose comme vraie, alors la connaissance à son sujet n'est pas entrée dans l'ontologie. Nous ne pouvons garantir la vérité (au sens ontologique, plus proche de l'existence) de ce qui n'est pas affirmé dans une ontologie. Dans la pratique, cela influence la manière dont les inférences sont produites par le moteur d'inférences puisqu'il est souvent nécessaire de « fermer » le monde par l'utilisation d'axiomes de clôture, c'est-à-dire avec des restrictions pour éviter d'avoir des déductions « aberrantes ».

La requête, quant à elle, est le corollaire de l'inférence. Une requête permet d'interroger une ontologie, selon les règles de logiques mises en place et l'obtention de réponses en fonction de l'axiomatisation de l'ontologie. Le langage le plus utilisé pour interroger les connaissances « en ontologies *OWL* » est le *SPARQL*[69] , mais il existe également le langage *SWRL*[70], qui permet d'exécuter des requêtes en langage *OWL* dans l'ontologie[71]. A partir de ce langage, des interfaces *endpoint* sont développées et permettent d'accéder à de la connaissance instanciée dans les ontologies. A l'heure actuelle, l'une des bases de connaissances construites sur des ontologies les plus en pointe est *DBpedia*[72], qui recense via des ontologies du Web sémantique les connaissances de l'encyclopédie collaborative *Wikipédia*. Nous ferons par la suite une rapide présentation de ce projet qui nous a inspirés pour nos propres travaux.

Inférences et requêtes reposent sur une taxinomie qui est organisée selon des règles définies à l'avance, *les axiomes*. Selon le CNRTL, un axiome[73] est un : « *Enoncé répondant à trois critères fondamentaux : être évident, non démontrable, universel. […]* ». Cette définition est étendue à un « *énoncé, proposition, posés à la base d'un système hypothético-déductif ou plus généralement élément d'une axiomatique*. Cf. loi logique, proposition logique a priori […]* ». Plus généralement, l'axiome est un « *énoncé admis comme base ou principe d'une construction scientifique […].* » et est utilisé couramment comme une « *vérité ou assertion admise par tous sans discussion […].* » Derrière l'ensemble de ces définitions, il s'agit d'exprimer l'absolue nécessité et la juste suffisance. L'axiome doit « tenir tout seul » ; il y a évidemment derrière ce concept le paradigme de vérité dans une forme d'absolu, mais dans notre cas, pour la construction d'ontologies, cela ne nous intéresse pas car il ne s'agit pas de définir le vrai du faux dans le « monde », mais de définir, dans l'ontologie, ce qui est considéré comme existant ou non. « Orange » peut être axiomatiquement défini comme « un fruit rond comestible », comme « une couleur qui signale un danger », ou comme « la couleur du fruit rond comestible » ou même finalement comme « un état d'esprit d'une personne sensible à la pluie » ; cette dernière proposition semble *a priori* « moins vraie » que les autres, mais elle peut tout à fait être un axiome de conception d'une ontologie pour laquelle il faudra accepter qu'elle le soit (il y a fort à parier cependant que l'engagement ontologique soit trop important pour pouvoir proposer

69 Source : (*SPARQL 1.1 Overview*, no date).
70 Source : (*SWRL: A Semantic Web Rule Language Combining OWL and RuleML*, no date).
71 Source : (O'Connor and Das, 2009).
72 Sources (*DBpedia Wiki | DBpedia*, no date; *DBpedia*, no date).
73 Définition donnée par le CNRTL (*AXIOME : Définition de AXIOME*, no date).

une ontologie qui soit universelle et perdre complètement l'ambition de représentation).

Il faut également comprendre que les axiomes font de l'ontologie un ensemble de déclarations à propos du monde que nous voulons représenter. Il s'agit donc de choix de représentations du monde ; c'est tout l'enjeu dans la conception de l'ontologie que nous retrouvons dans sa construction. Quoi qu'il en soit, les axiomes donnent corps à l'ontologie, elle n'est plus seulement concept, elle devient structure sémiotique[74].

Nous avons construit très rapidement une ontologie de domaine « navires ». Elle a la prétention de respecter les critères de conception de Gruber et Guarino. Notamment, le principe de subsomption est respecté entre les classes et nous avons cherché à minimiser notre engagement ontologique dans l'axiomatisation. Elle n'a aucune autre prétention que de servir à l'illustration des concepts présentés.

Voici un exemple de ce que peut être un ensemble d'axiomes dans cette ontologie :

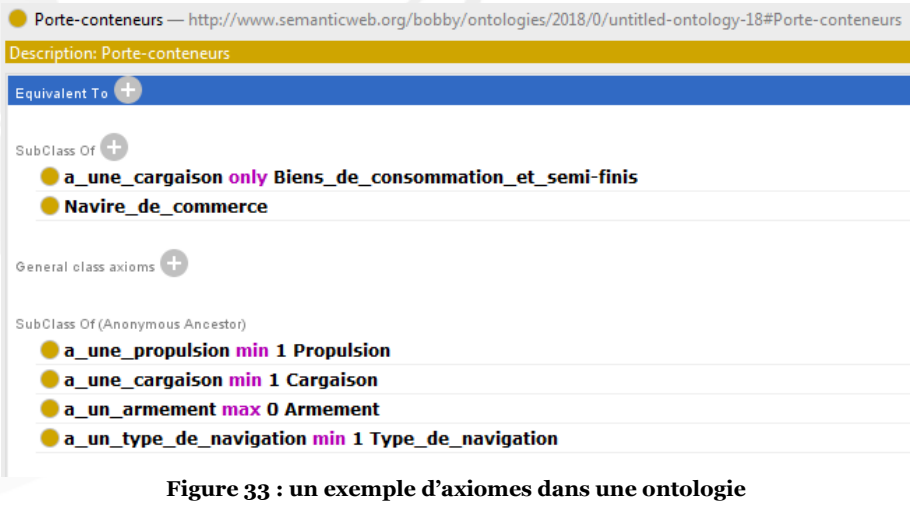


Figure 33 : un exemple d'axiomes dans une ontologie

De façon déclarative, nous définissons *Porte-conteneurs* comme une classe définie selon six axiomes. Les axiomes sont construits selon un modèle *sujet, prédicat, objet* pour être pleinement opérés en logique des prédicats. Dans notre cas, la classe *Porte-conteneurs* est définie comme un navire de commerce qui transporte seulement (le cas échéant) des biens de consommations et semi-finis. Il a au moins une propulsion, au moins une cargaison, aucun armement et au moins un type de navigation.[75].

Certains axiomes sont spécifiques à la classe *Porte-conteneurs*, d'autres sont hérités des classes supérieures, c'est une propriété inhérente à la taxinomie. C'est ce dont nous discutons maintenant.

74 De la définition 2.A.1 du CNRTL (*SÉMIOTIQUE : Définition de SÉMIOTIQUE*, no date).
75 Les termes en rose sont des restrictions ou des cardinalités selon le langage *OWL* : c'est une évolution importante du standard RDF qui permet de contraindre à une existence ou une valeur des axiomes et d'amener des possibilités de représentation du monde plus importantes.

2.2.2.2 Taxinomie

Les axiomes une fois définis vont permettre de créer *une taxinomie*. La taxinomie est la « science de la classification »[76]. Classifier, c'est créer une hiérarchie de classes, « méthodique ». Une manière de le faire est de suivre *le principe de subsomption*[77], de « *[…] penser le particulier sous le général […]* ». Par exemple :

$$LNG\ tanker \sqsubseteq navire\ citerne \sqsubseteq navire\ de\ commerce \sqsubseteq navire$$

LNG tanker possède les propriétés de *navire-citerne* qui possède celles de *navire de commerce* qui possède celles de *navire*. En logique ensembliste, on l'illustrerait de la sorte :

$$E_{LNG\ tanker} \subset E_{navire-citerne} \subset E_{navire\ de\ commerce} \subset E_{navire}$$

Un individu qui appartient à une classe possède donc les propriétés de cette classe et hérite des propriétés des classes supérieures. En continuant de la sorte, on crée un arbre des similarités et des différences en fonction des axiomes qui définissent l'arborescence. On peut donc classer les individus. Si l'arborescence est la suivante :

$$E_{réserve} \subset E_{officier} \subset E_{marine} \subset E_{armée}$$

et un individu tel que :

$$x_{Thibaut} \in E_{réserve}$$

Alors on pourrait en déduire seulement et suffisamment (ni plus ni moins) que Thibaut est officier (de réserve) dans la marine de guerre. Cependant, on voit que cette très « légère » arborescence a de nombreux défauts de conception dans sa construction, en particulier :

L'ensemble « réserve » est inclus dans l'ensemble « officier » ; cela implique qu'il n'est pas possible d'être « dans la réserve » sans être officier ; ce qui est faux dans au moins un monde, « l'armée française en 2018 », le principe de subsomption n'est pas vérifié. Il s'agit ici d'un biais volontaire ou non de la représentativité de l'ontologie.

Plus généralement, la cohérence de l'arborescence est discutable ; en effet, autant l'arborescence « navire » présentée auparavant respecte le principe de subsomption, autant, ici, ce n'est pas explicite ; l'ensemble « marine » pourrait tout à fait être inclus dans un autre ensemble que « armée » et les propriétés inhérentes de « armée » n'ont *a priori* rien d'impératif pour concevoir l'ensemble « marine ». Pour le coup, il va donc falloir poser des axiomes particulièrement explicites pour montrer en quoi ici l'ensemble « marine » est inclus dans l'ensemble « armée ». Ce n'est pas une erreur en soi, mais on s'éloigne de l'objectif d'explicitation (et notamment du critère « *biais cognitif minimum* ») proposé par Gruber pour la conception ontologique. Guarino serait encore plus critique et ne pourrait considérer cette taxinomie comme structure pour une ontologie.

---

76 Définition par extension donnée par le CNRTL (*TAXINOMIE : Définition de TAXINOMIE*, no date).

77 Définition issue du CNRTL (*SUBSOMPTION : Définition de SUBSOMPTION*, no date). Voir la partie *OntoClean*.

Plus généralement, il y a une « bonne dose » d'implicite dans cette arborescence et il semble qu'il manque des classes intermédiaires (et la construction axiomatique afférente) pour comprendre au mieux le « passage » d'une classe à l'autre.

Ces deux exemples montrent rapidement l'importance à accorder à la définition des axiomes qui vont déterminer l'arborescence des ensembles (la hiérarchie des classes dans le langage ontologique).

Nous allons voir maintenant que la définition des axiomes est d'autant plus critique qu'elle détermine également les relations entre les classes ; ces relations qui représentent les associations des classes entre « au travers » de la hiérarchie. Ces relations sont le pendant de la hiérarchie pour la représentation du monde dans l'ontologie. Pour revenir à notre exemple de navire, la classe *Porte-conteneurs* se situe dans la hiérarchie suivante :

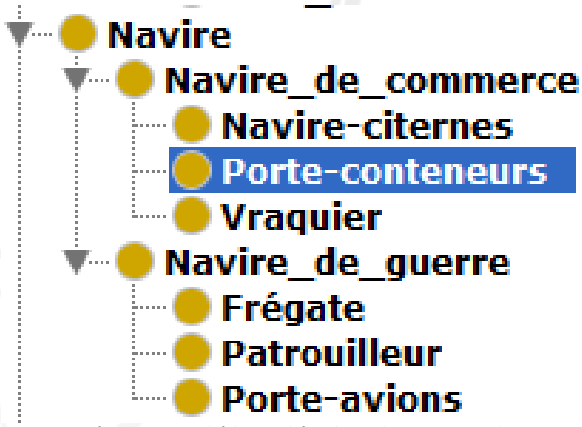


Figure 34 : hiérarchie des classes *Navire*

On en déduit que la classe *Porte-conteneurs* est une sous-classe de *Navire de commerce*, elle-même une sous-classe de *Navire*. En conséquence de cela, la classe *Porte-conteneurs* hérite des propriétés des classes supérieures donc, dans notre cas, les axiomes qui définissent *Navire* et *Navire de commerce* définissent aussi *Porte-conteneurs*. Ce sont les axiomes en héritage (ceux qui maintiennent l'appartenance) :

Figure 35 : axiomes en héritage pour la classe *Porte-Conteneurs*

Enfin, on a les axiomes spécifiques à la classe (ceux qui créent la différence) :

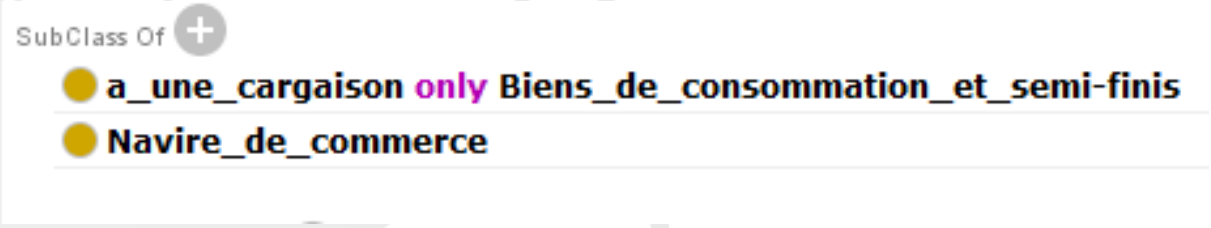


Figure 36 : axiomes spécifiques de la classe *Porte-Conteneurs*

Les axiomes forment le « squelette sémantique » de l'ontologie construit selon le modèle *sujet, prédicat, objet*. Ils sont opérables selon la logique des prédicats pour apporter des connaissances déduites grâce aux inférences ou recherchées grâce aux requêtes.

2.2.3 Le modèle de description de ressources dit standard *RDF*

Nous allons maintenant nous pencher sur un modèle proposé par l'ingénierie des connaissances, à la fois conceptuel et structurel, qui permet une implémentation formelle, logique et informatisée de l'ontologie. Cela autorise à construire « concrètement » une ontologie. Il s'agit du *modèle de description des ressources,* ou le *Resource Description Framework (RDF)*. C'est un langage (devenu standard) pour représenter l'information à propos des ressources accessibles sur Web.

2.2.3.1 Les schémas RDF, RDFS Semantics

Le Web fournit une forme générale d'identificateur de ressources : l'*URI*, *Uniform Resource Identifier*. Une *URI* peut être créée pour faire référence à tout ce qui doit être mentionné dans une déclaration (« *statement* »). Tout peut donc être identifié de la sorte, objets mentaux ou physiques, ressources matérielles, êtres vivants, organisations, absolument tout.

Le *RDF* utilise les *URI*[78] comme les bases de son mécanisme d'identification des sujets, des prédicats et des objets dans les énoncés. La structure de base du *RDF* est à trois composants toujours écrit dans l'ordre sujet, prédicat, objet[79] (« RDF Primer, » n.d., chap. 2.1 Basic Concepts) :

- le sujet, identifie la chose dont il est question dans l'énoncé (dans la proposition) ;
- le prédicat, identifie la propriété ou la caractéristique du sujet que l'énoncé spécifie ;
- l'objet, identifie la valeur de cette propriété.

On retrouve évidemment la logique $P(s, o)$ présentée auparavant. Le prédicat est également connu sous le nom de propriété du *triple*. Il indique une relation. Chaque

78 Il y a également une volonté de normalisation de ces URIs par la création d'IRI, *Internationalized Resource Identifier* qui répondent à la norme présentée dans le document RFC 3987 (*RDF 1.1 Concepts and Abstract Syntax*, no date, chap. 3.2 IRIs).
79 La logique afférente à ce modèle est présentée dans la suite du document, nous ne montrons ici que la structure : c'est un modèle assez intuitif finalement.

triple représente un énoncé d'une relation entre les choses indiquées par les nœuds qu'il relie.

Le graphe est l'illustration d'une classe (*set*) de *triples :*

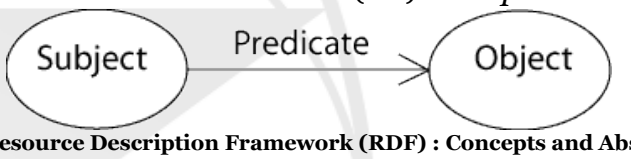


Figure 37 : *triple* (« Resource Description Framework (RDF) : Concepts and Abstract Syntax, » n.d.)

Le sens de l'arc est significatif : il pointe toujours vers l'objet. C'est un graphe orienté. Les classes des nœuds d'un graphe *RDF* sont les classes des sujets et objets des *triples*. A partir de cette structure, un ensemble de classes et de propriétés est décliné. Cet ensemble constitue un schéma du standard *RDF*.

Le Web sémantique a proposé en 2004 un modèle de description des ressources (ici des documents, des pages Web) construit sur un standard que nous allons présenter qui est devenu la base de l'immense majorité des ontologies informatisées à l'heure actuelle (« Le W3C publie les recommandations RDF et *OWL*, » n.d.).

Le schéma du standard *RDF* à jour à l'heure actuelle (en janvier 2018) est la version 1.1[80]. Il s'agit en fait du « méta-modèle » du *RDF* dans le sens où le schéma *RDF* fournit un vocabulaire de modélisation de données pour le modèle *RDF*. En particulier, le schéma *RDFS* ou *RDF Semantics* qui formalise la formalisation (« RDF Semantics, » n.d., chap. 0.1 Specifying a formal semantics : scope and limitations)[81]. On amène un langage formel à la manière de décrire et de relier les entités entre elles avec l'un des objectifs qui est de respecter ce que l'on pourrait appeler en français le « sous-entendu », dans le sens « qui implique », « qui aide à comprendre » tel que :

> « *Une relation sémantique entre les expressions qui tient chaque fois que la vérité de la première garantit la vérité de la seconde. De façon équivalente, chaque fois qu'il est logiquement impossible que la première expression soit vraie et la seconde fausse. De la même manière, lorsque toute interprétation qui satisfait la première satisfait également la seconde. (Également utilisé entre un ensemble d'expressions et une expression).* » (D. Glossary of Terms (Informative) « RDF 1.1 Semantics, » n.d.)

Cette relation sémantique est une articulation fondamentale en logique, mais elle n'est pas toujours explicite dans le langage naturel et il y a donc nécessité d'apporter un formalisme logique à l'expression de la langue lorsque l'on souhaite construire une ontologie ; cela rejoint la volonté d'éliminer toutes les ambiguïtés.

La version la plus à jour de la description du schéma *RDFS* est présentée dans le document (« RDF 1.1 Semantics, » n.d.). Ce schéma sert de fondation pour le langage *OWL*.

80 Sur le site du W3C, (*RDF Schema 1.1*, no date).

81 Notons que le terme « *semantics* » utilisé ici correspond seulement à un modèle de description des données (« *model theory* » associant les expressions aux interprétations) et n'a pas de lien direct avec le travail que nous menons dans ce domaine si ce n'est une quête de la signification.

2.2.3.2 Le langage OWL

Le consortium W3C a mis au point le langage *OWL*, pseudo-acronyme pour *Web Ontology Language*, un langage du Web sémantique pour « représenter des connaissances riches et complexes sur les choses, les groupes de choses et les relations entre les choses. *OWL* est un langage basé sur la logique computationnelle tel que les connaissances exprimées dans *OWL* peuvent être exploitées par des programmes informatiques, par exemple pour vérifier la cohérence de ces connaissances ou pour rendre explicites les connaissances implicites. » (« *OWL* – Semantic Web Standards, » n.d.). « Les documents *OWL*, [sont] connus sous le nom d'ontologies. […] »[82]. Le langage *OWL* a connu une évolution et la dernière version est la version *OWL* 2[83]. Le langage *OWL* emprunte de nombreuses propriétés au standard RDF qu'il enrichit en fournissant « un vocabulaire supplémentaire et une sémantique formelle »[84].

Le langage *OWL* ne modélise pas (seulement) une description des ressources (*data*), mais des connaissances (knowledge) : il est le langage informatique des ontologies. Dans le document (« *OWL* 2 Web Ontology Language Primer (Second Edition), » n.d., chap. 3 Modeling Knowledge : Basic Notions), on trouve la description du modèle des connaissances :

- axiomes, ou les déclarations de base qu'une ontologie *OWL* exprime ;
- entités, ou éléments utilisés pour se référer à des objets du monde réel ;
- expressions , ou combinaisons d'entités pour former des descriptions complexes à partir de descriptions de base.

Cette description des connaissances est complètement en phase avec la logique des prédicats présentée auparavant. Le meilleur manuel d'explication de la construction ontologique avec l'utilisation du langage *OWL* est celui de Horridge et al. (2011). Signalons aussi le papier de Raimbault et al. (2008) qui proposent un panorama de la « galaxie *OWL* » au travers des graphes conceptuels et la présentation à propos du langage *OWL* de Lacot[85]. Regardons maintenant un exemple très simple[86] :

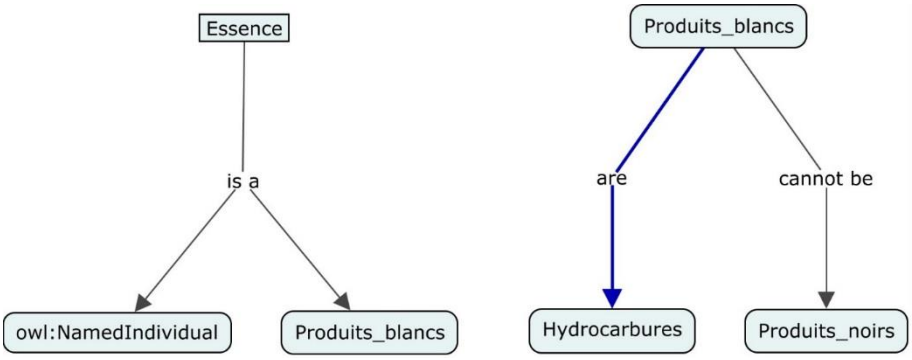


Figure 38 : schéma d'une description écrite en langage *OWL* d'un hydrocarbure

82 Source : (*OWL - Semantic Web Standards*, no date).
83 Source : (*OWL 2 Web Ontology Language Document Overview (Second Edition)*, 2012).
84 Source : (*OWL Web Ontology Language Overview*, no date).
85 Source : (JoliCode, no date).
86 Ce schéma a été effectué à l'aide du logiciel (*CmapTools Ontology Editor*, no date) version adaptée pour les ontologies de CmapTools (*Cmap | Cmap Software*, no date) nous y reviendrons par la suite.

Les instances sont les formes rectangulaires aux bord droits tandis que les classes sont les formes aux bords arrondis. Les prédicats sont représentés par les relations orientées d'une forme à une autre. Le schéma se lit de gauche à droite (en l'occurrence) :

- *Essence* est un *Individu*, une instance de la classe *Produits blancs ;*
- la classe *Produits blancs* est une sous-classe de la classe *Hydrocarbures* (par subsomption) ;
- la classe *Produits blancs* est disjointe de la classe *Produits noirs* : un individu ne peut appartenir aux deux classes simultanément.

Si on reprend l'ontologie « navires » qui nous a servis d'exemple précédemment, on peut définir ainsi la classe *Porte-conteneurs* :

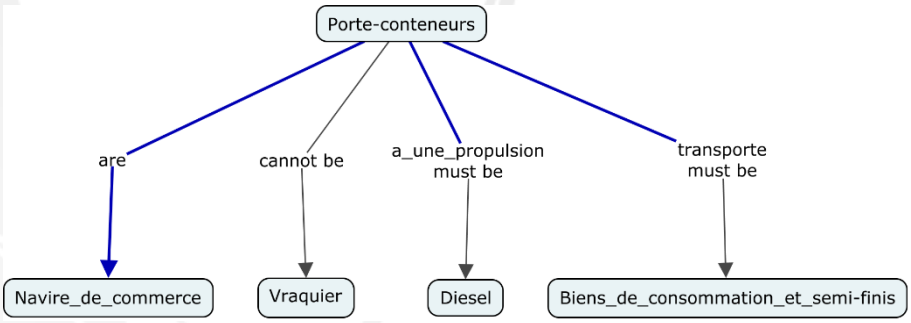


Figure 39 : le graphe *OWL* de la classe Porte-conteneurs

Et une *Instance* de *Porte-conteneurs*, le navire Maersk Garonne[87] :

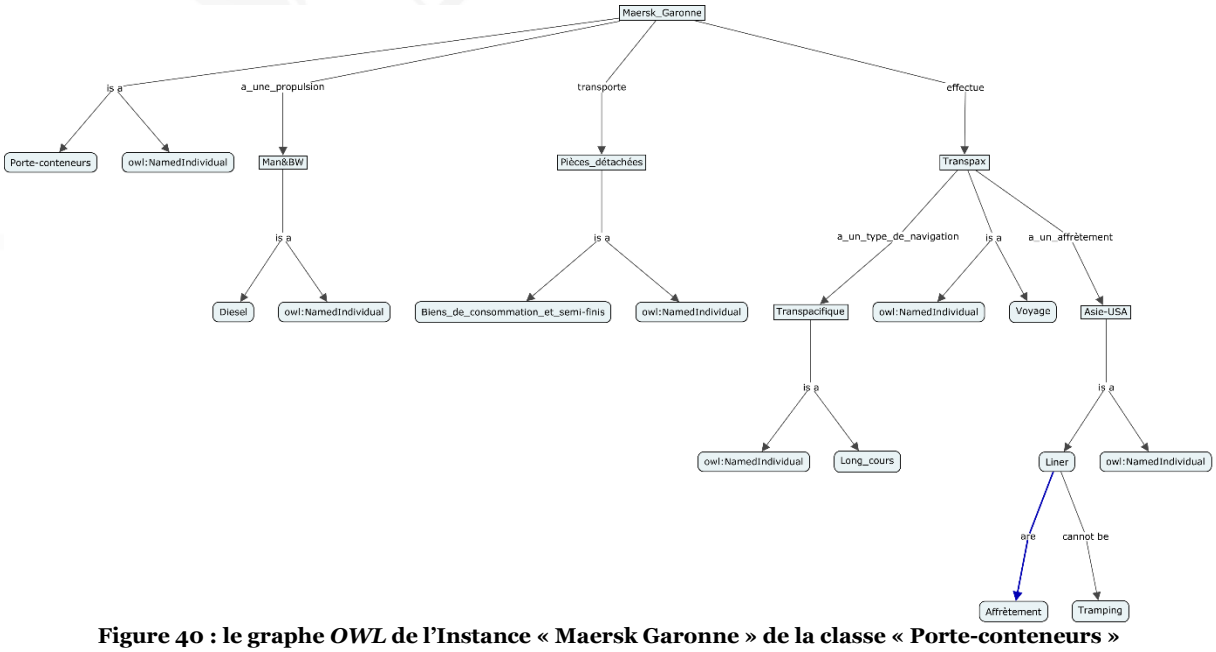


Figure 40 : le graphe *OWL* de l'Instance « Maersk Garonne » de la classe « Porte-conteneurs »

Le langage *OWL* permet la description du « monde » dont l'ontologie est la représentation. Il est, à notre connaissance, un des langages les plus utilisés et maintenus. Il faut signaler qu'il a comme « concurrent » principal le langage *OBO,*

87 Navire français magnifique sur lequel l'auteur de cette thèse a fait ses premières armes en tant qu'officier (*Vessel details for: MAERSK GARONNE (Container Ship) - IMO 9235579, MMSI 219219000, Call Sign OUTK2 Registered in Denmark | AIS Marine Traffic*, no date).

Open Biological and Biomedical Ontology, bien qu'il semble exclusivement utilisé dans les ontologies de domaines en biologie et génétique (« The OBO Foundry, » n.d.).

Nous venons de voir une présentation du langage *OWL* qui permet une implémentation informatique efficace pour la réalisation d'ontologies scientifiques. Cela sera le langage que nous utiliserons dans la suite de nos travaux et nous nous tiendrons à sa syntaxe et sa sémantique pour la description de ce que nous faisons.

2.3 Choix d'une ontologie

Dans cette section, nous présentons l'ontologie que nous avons retenue pour nos travaux sur l'accident de *Deepwater Horizon*, ainsi que le cheminement qui nous y a conduits. Nous débutons par un tour d'horizon de quelques réalisations d'ontologies susceptibles de nous inspirer ou vers lesquelles il était naturel de se tourner en premier lieu.

2.3.1 Tour d'horizon de quelques réalisations

Le laboratoire dans lequel travaille l'auteur de cette thèse est à l'origine d'une contribution importante en matière de conception et de déploiement pleinement opérationnel d'une ontologie, juridique, à la base d'un logiciel de conformité réglementaire très répandu (Vigneron, Rallo and Guarnieri, 2014).

Mais dans notre domaine, le constat à propos de l'existant en matière de bases de données d'accidents accessibles est assez sévère, en tout cas en France. Par exemple, le site institutionnel de référence à propos des accidents industriels en France est le site Aria[88] ; lorsqu'on y recherche des informations à propos du cas *Deepwater Horizon* avec la requête « *Deepwater Horizon* », le site propose trois résultats qui ne correspondent pas à ce cas[89]. Il nous a fallu procéder par recoupements successifs pour retrouver le cas *Deepwater Horizon*. En mettant de côté le contenu qui renferme de nombreuses erreurs et approximations, le plus surprenant est l'absence d'un formalisme de description et une identification défaillante des sources.

Le CEDRE, CEntre de Documentation, de Recherche et d'Expérimentations sur les pollutions accidentelles des eaux, est nettement plus au point sur l'accident, mais surtout sur la structuration de l'information disponible[90], cependant, le traitement du cas reste limité.

L'existant en matière de bases de connaissances n'est pas du tout à la hauteur de l'évènement que nous avons déjà décrit en partie dans le chapitre 1.

D'une manière générale, l'état de formalisation des connaissances institutionnelles sur le cas *Deepwater Horizon* s'avère bien en deçà de ce que nous sommes en droit d'attendre à propos d'un tel évènement et il y a matière à produire

88 Sources :https://www.aria.developpement-durable.gouv.fr/, https://www.aria.developpement-durable.gouv.fr/accident/38145/. Et les trois mauvais résultats (en septembre 2018) : https://www.aria.developpement-durable.gouv.fr/?s=deepwater%20horizon&fwp_sort=date_desc.

89 Nous avons essayé avec Macondo comme mot-clé, mais cela ne donnait rien non plus.

90 http://wwz.cedre.fr/, http://wwz.cedre.fr/Ressources/Accidentologie/Accidents/Deepwater-Horizon

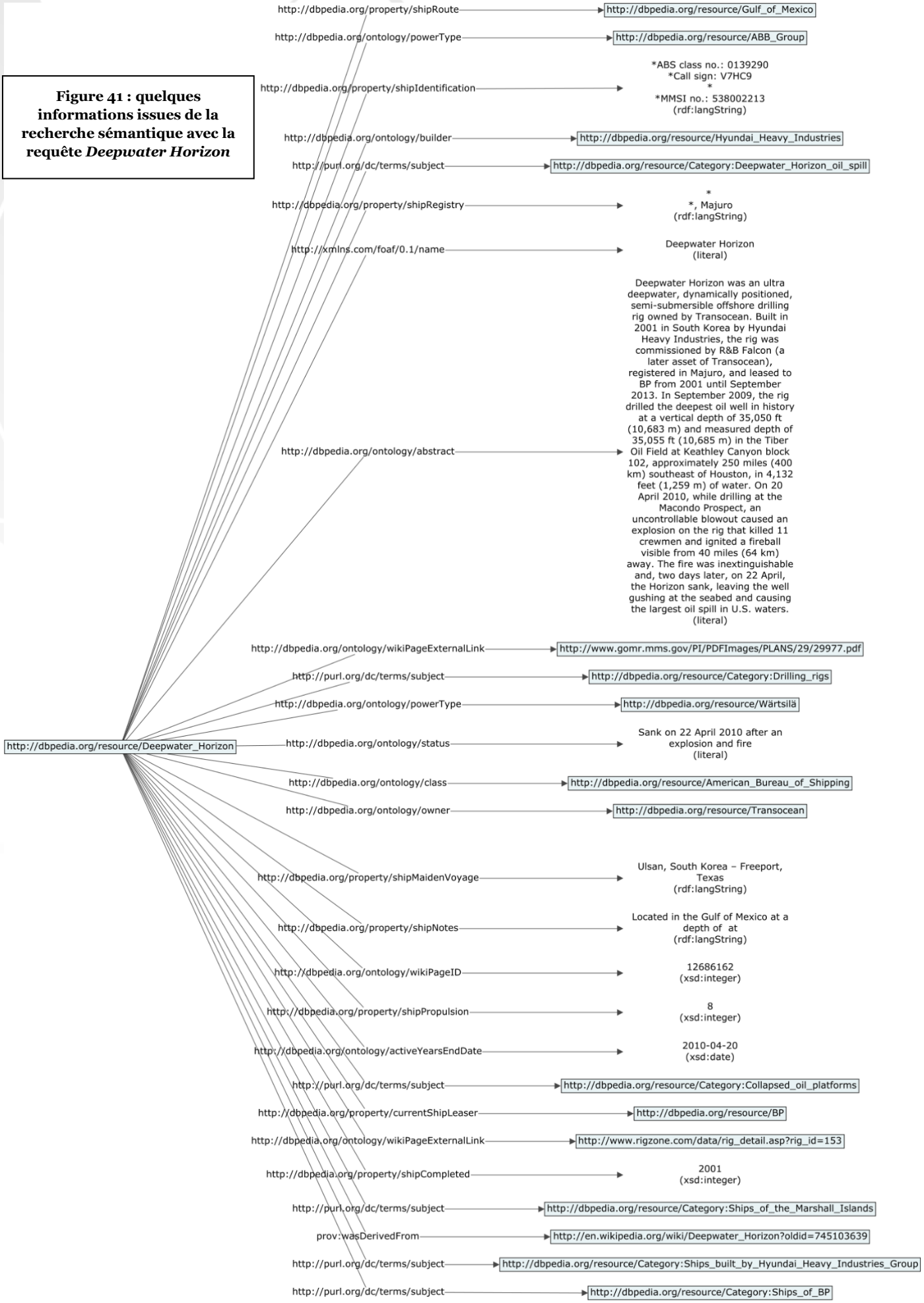
quelque chose de nettement meilleure. Le projet *DBpedia*, très rapidement évoqué auparavant, nous a inspirés pour essayer de faire de même pour le cas *Deepwater Horizon*.

DBpedia[91] est un projet communautaire à grande échelle d'extraction de données pour les structurer et créer un réseau de connaissances sur l'immensité des informations écrites dans l'encyclopédie en ligne *Wikipédia*. Google utilise une approche similaire pour créer ces fiches de connaissances dans les résultats de recherche. Le projet a donc pour objectif d'étendre « la connaissance » disponible[92] dans *Wikipédia* par la construction de liens sémantiques entre la pléthore des instances (au sens atomique) de savoirs agrégés via le site. A l'heure actuelle, le projet a recensé et décrit en langue anglaise 4,58 millions d'entités (personnes, évènements, lieux, travaux et créations, phénomènes…) dont 4,22 millions de ces entités ont été classifiées dans une ontologie consistante. Il ne s'agit pas de juger de la qualité de l'information disponible dans *Wikipédia* ou de la pertinence des liens sémantiques (très discutables pour la requête que nous allons effectuer), mais simplement de montrer l'intérêt en matière d'enrichissement des connaissances par la liaison sémantique. Pour avoir la vision complète des résultats, il faut passer par une interface spécifique comme un navigateur sémantique (*faceted browser)* ou un logiciel comme *Protégé* pour « naviguer sur le graphe des données ». Nous présentons ici quelques résultats à valeur d'illustration pour la requête « *Deepwater Horizon* «. Nous avons adopté le logiciel *COE CmapTools*, le plus performant selon nos tests pour manipuler, éditer et représenter des ontologies scientifiques codés en *OWL*. Le graphique ci-après a été épuré de nombreux éléments pour rester lisible dans l'espace d'une feuille A4.

Chaque relation est construite sur le modèle du *triple* et de nombreuses ontologies de description des ressources qui traitent sous différentes facettes la requête sont sollicitées pour établir ce graphique. Cette représentation graphique est le pendant formalisé de notre « modèle stratosphérique » que nous avions intuitivement construit[93] au début de nos travaux de recherche.

91 Source : (*DBpedia*, no date).
92 Nous parlerions plus volontiers d'information, mais « connaissance » est le terme consacré dans ce projet.
93 Voir le chapitre 1.

Figure 41 : quelques informations issues de la recherche sémantique avec la requête *Deepwater Horizon*



Fort de notre apprentissage en matière de construction sémantique en lien avec le projet *DBpedia*, et afin de montrer des pistes d'amélioration de la connaissance institutionnelle sur le cas *Deepwater Horizon*, nous avons élaboré un récit de l'accident en utilisant le moteur de gestion de contenu *MediaWiki*[94], renforcé par l'extension *SemanticMediaWiki*[95], dans le but de produire une description sémantique de l'accident à partir de nos propres données et de notre approche experte du domaine. Il n'est pas possible de montrer dans ce manuscrit l'entièreté de ce récit, mais nous en proposons ci-après un extrait :

The accident

The Deepwater Horizon accident has no precedent in the offshore industry history. This is the greatest environmental disaster of the US history. Rig fires and explosions often happen on the GoM area with more than an hundred a year since 2007 with no real trend of improvement. Fortunately, most of these accidents are without heavy consequences but it is a likely risk. In 1979, the Ixtoc I rig encountered a subsea blow-out that led to the release of a very important amount of oil, but without any official investigation neither report nor trial facts, it is difficult to clearly estimate what was the quantity of oil released and the global cost of the response and restoration. Even if it could seem similar, there were major differences that clearly separate the two disasters. The wellhead of Ixtoc I was only 160 ft. deep compared to the 5067 ft. of the Deepwater Horizon: the access requests high technology subsea systems and the behavior of an oil plume is completely unknown at this depth. Moreover, inspection and emergency works or interventions have never been performed before. The technology of drilling has been greatly improved and the potentiality of a heavy release of hydrocarbons has increased during the last 30 years, the oil spill response technology, lesser less. The apparition of the Internet and social networks and the correlated public scrutiny. The growth of the public interest and the civil society on the responsibilities of the companies managing their business and particularly the oil&gas companies. The geology of the Block 252 is not well known and it will be the first well drilled by BP in these lease area. As several others wells in the area, this was considered by the drillers themselves as "a well from the hell", one of the very deep well drilled in the GoM (planned depth was of 22000ft). The Exxon Valdez experience feedback, including oil spill response sizing and planned emergency actions, was strongly not sufficient. The total cost of the response including criminal fines, today, is estimated as far as 62 $ billion.

Figure 42 : extrait de texte concernant l'accident de *Deepwater Horizon*

Dans ce récit, nous créons des liens sémantiques qui nous paraissent pertinents et qui structurent le récit pour qu'il puisse être opérable par la machine (standard *RDF*). Ci-dessous, nous voyons les prédicats créés « dans l'extrait » ci-dessus et nous revenons à notre fil conducteur que nous avons découvert avec le Web sémantique, soit la possibilité d'apporter directement la réponse à un questionnement.
Par exemple, nous pouvons répondre à la question « *Where does the Deepwater Horizon accident occurred ?* » car nous avons structuré au préalable une réponse autour du prédicat *Has a location*.

Browse wiki

Main Page

Has a location	50 Nm (South, off the Louisiana Coast USA) +
Has an emergency situation	The evacuation of the rig +
Has casualties	11 person* +
Has dangerous phenomenom	A first explosion at 21:49 and a second one 10 sec after. +
Has date	April 20, 2010 +
Has hazardous conditions	Gas-air mix created an explosive atmosphere +
Has initial crew	126 person +
Has page	Deepwater Horizon drilling rig + , The Blow-Out Preventer + , Kick + , United States Coast Guard (USCG) + , Drill a relief well + , Search and Rescue (SAR) operation + , The discovery of the leaks + , A well control situation + , The National Contingency Plan (NCP) + , Incident Command Post (ICP) + , The growing concern of the pollution risk + , Engineering approach + , The National Incident Commander (NIC) + , The cofferdam + , The Containment and Disposal Project + , The Top Hat 4 + , The 3-ram capping stack + , The Riser Insertion Tube Tool (RITT) + , The explosions and the fire + , The beginning of the SAR party and the emergency intervention on the BOP + , The official kill declaration of the Macondo well + , Spill Of National Significance + , The oil spill response + , The worst environmental disaster of the US history + , The Macondo well + , The accident +
Modification date	19:59:16, 29 January 2018 +

hide properties that link here

No properties link to this page.

Retrieved from "http://localhost/mywiki/index.php/Special:Browse/:Main-5FPage"

Figure 43 : les prédicats sémantiques associés à l'extrait en exemple

94 Source : (*MediaWiki/fr - MediaWiki*, no date).
95 Source : (*Présentation de Semantic MediaWiki — semantic-mediawiki.org*, no date).

On peut aussi lier des données avec d'autres permettant un référencement ou un *sourcing* efficace :

Figure 44 : exemple de lien d'une donnée vers un site web

Nous répondons à la question de la localisation, en donnant le lieu géographique de l'évènement, et nous invitons le lecteur (ou la machine) à aller « voir » vers le lien en question s'il souhaite confirmer sa lecture ou en apprendre davantage. Les possibilités sont très importantes et nous balbutions pour le moment. Cette approche nous paraît très intéressante et nous aimerions créer une monographie de cet accident de la sorte. L'intérêt majeur serait de fournir par la suite une base de connaissances structurée à l'aide du réseau sémantique que l'on pourrait enrichir, et qui servirait de référence en matière de connaissances pour un cas d'accident.

A propos de *DBpedia*, la structuration apportée par le langage *RDF* est utilisée à la fois comme descriptrice de la nature des données (des fragments de textes), *de facto* des métadonnées, mais aussi comme descriptrice sémantique de ce que représentent ces données, ce qui met « au même niveau » deux représentations différentes des connaissances et qui ne nous paraît pas pour autant pertinent[96]. D'ailleurs, *DBpedia* structure plutôt ce que nous avons appelé « information » à la lumière du chapitre 1 tandis que notre proposition de récit sémantique se rapproche plus d'une connaissance construite par la compréhension des informations que nous possédons (nonobstant la proposition de *DBpedia*). Or, nous souhaitons apporter une représentation des connaissances à propos de notre objet d'étude, et donc aller au-delà d'un recensement, même bien qualifié, de relations entre informations.

Enfin, comme nous pouvons le voir avec notre exemple, l'approche est *ad hoc* (surtout à notre niveau, mais également dans une moindre mesure avec *DBpedia*), notamment dans la création des prédicats et donc notre travail reste très (trop) dépendant du domaine. L'ontologie, utilisée dans *DBpedia* est une ontologie au « sens faible ».

Pour ces raisons, mais également pour des raisons qui paraîtront plus clairement dans le chapitre 3, il nous est apparu fondamental de partir à la recherche d'une ontologie « idéale ».

96 Voir la page en question comme exemple : http://DBpedia.org/page/Deepwater_Horizon_oil_spill.

2.3.2 La recherche de l'ontologie « idéale »

Nous avons vu comment les ontologies étaient conçues, construites et utilisées dans le cadre de l'ingénierie des connaissances et ce qu'elles pouvaient nous apporter pour résoudre notre triple problème de données, de nature et de représentation du cas *Deepwater Horizon*. Nous avons cerné en quoi une ontologie de haut niveau pourrait être la solution la mieux adaptée et nous allons maintenant nous mettre en chasse de l'ontologie qui correspondra le mieux à nos attentes pour résoudre notre triple problème. Nous avons élaboré notre propre cahier des charges, à savoir :

- être peu coûteuse en ressources cognitives, « facile » à appréhender par un non-spécialiste ;
- être capable de représenter de multiples réalités plus ou moins partagées par une multitude d'acteurs, en très grande quantité ; avoir la possibilité d'affirmer le caractère socialement construit des faits et des controverses ;
- être capable de suivre « à la trace » la donnée ; c'est-à-dire proposer à l'utilisateur la possibilité de connaître l'origine de l'information que le « fournisseur » amène à l'ontologie ;
- avoir une « réelle » portée universelle, c'est-à-dire, être accessible facilement en langage naturel afin de permettre à n'importe qui de la comprendre et de la « peupler » et avoir l'engagement ontologique le plus faible possible car nous l'avons vu, cela permet *de facto* d'avoir une portée, ontologique justement, de très grande ampleur et d'éviter une idiosyncrasie très préjudiciable ;
- se garder la possibilité, malgré tout, de pouvoir modifier l'ontologie, au niveau des classes et/ou des prédicats, donc de pouvoir « spécifier » l'ontologie de façon précise et formelle ;
- être lisible et « questionnable » par la machine, c'est-à-dire être formalisée par un langage informatique existant et si possible adapté aux besoins et outils actuels ;
- être exportable, partageable et utilisable sur la majorité des plateformes informatiques.

Suivant ces critères, nous avons usé de différentes approches dans notre revue de littérature pour trouver l'ontologie qui nous paraîtrait la mieux adaptée à notre besoin de représentation. Nous précisons d'emblée que nous n'avons pas envisagé au départ de créer notre propre ontologie, gageant qu'il en existerait une satisfaisante et que nous la réutiliserons ; nous ne réinventerons pas la roue.

2.3.2.1 Recherche par domaine

Dans une première phase, nous avons recherché des ontologies qui traitent d'accident de quelque manière que ce soit. Quelques auteurs ont produit des ontologies dans le domaine des « catastrophes, accidents, urgences » comme Babitski et al. (2009) , Batres et al. (2014), Delir Haghighi et al. (2013) et Xu et Zlatanova (2007). Mais ces ontologies sont des ontologies d'applications et sont clairement axées sur la

partie opérationnelle de la gestion d'accidents et d'urgences dans un but d'améliorer l'efficacité des services d'urgences notamment. Elles servent en fait plus de structures pour des bases de données et des systèmes experts qui permettent aux moyens de secours de se renseigner et de s'adapter en fonction des situations rencontrées. Ces travaux ne décrivent pas un accident en particulier, ce ne sont pas des représentations de l'accident.

Les travaux de Provitolo et al. (2009) présentent une ontologie de domaine des risques et catastrophes à un niveau conceptuel et cherchent à formaliser un langage partagé pour l'accident et la catastrophe. Nous avons trouvé l'approche trop hermétique et sans visibilité pour l'utilisation que nous souhaitons.

Les ontologies dans le domaine « catastrophes, accidents, urgences » que nous avons consultées nous donnent une impression de « figer » l'objet d'étude dans un statisme de faits et de temporalités. On fixe les choses d'une seule manière possible dans le « formol » et on accumule les autres cas de la même manière et au même endroit.

Nous avons alors élargi notre recherche au secteur d'activité dans lequel s'est produit l'accident de *Deepwater Horizon*. Nous avons trouvé des travaux intéressants sur une ontologie de la « marée noire », mais qui ne semblent malheureusement plus d'actualité[97]. Cependant, la portée de ces travaux n'est pas suffisante pour la considération que nous avons de notre cas et nous ne pouvons pas nous limiter à cette représentation. Comme nous l'avons déjà souligné, l'ontologie à la source de l'ISO 15926 n'est pas adaptée à notre propos.

En résumé, notre recherche d'une ontologie appliquée aux domaines connexes à l'industrie pétrochimique ou aux accidents industriels s'est avérée infructueuse.

2.3.2.2 L'approche paradigmatique, l'accident comme un évènement

Nous sommes alors revenus aux quatre piliers d'une ontologie, pour nous demander de quelle manière nous conceptualiserions une ontologie correspondant à nos besoins.

Selon O_1, nous voulons une ontologie avec un engagement le plus faible possible, probablement une ontologie de haut niveau, mais surtout une ontologie indépendante du domaine.

Selon O_2, nous cherchons à représenter le cas *Deepwater Horizon* parce qu'il ne « rentre » pas dans les modèles d'accident existants et que la science l'a plus ou moins mis de côté, particulièrement la période de 87 jours d'intervention que nous n'avons vue traitée quasiment nulle part. Ce cas nous est mis à disposition très majoritairement au travers de documents écrits. Il associe des existants humains et des non-humains, et des phénomènes naturels.

Selon O_3, l'ontologie devra pouvoir faire s'exprimer les caractéristiques générales de cet évènement, les mécanismes causaux à l'œuvre notamment, tels qu'ils sont décrits par les documents écrits, parfois contradictoires.

97 Source : https://cordis.europa.eu/project/rcn/93797_en.html

Selon O_4 enfin, on doit pouvoir mettre en lumière les zones de connaissances consolidées, et celles plus fragiles. L'ontologie est ici vue comme un moyen d'orienter des efforts de recherche pour élucider des mécanismes causaux insuffisamment compris ou documentés.

A partir de cette mise à plat de notre réflexion, il n'y a que très peu de papiers qui ont retenu notre attention : citons les travaux de Kaneiwa et al. (2007), Scherp et al. (2009) et Shaw et al. (2009) qui proposent des ontologies de haut niveau qui sont *a priori* intéressantes par leur engagement ontologique à propos de l'évènement ; mais leur point commun nous a interpellé : elles sont toutes construites ou inspirées d'une ontologie de haut niveau, *DOLCE*. *DOLCE* est une ontologie « racine » qui est déclinée en versions dont une, *DnS UL* dite *DUL* qui est une ontologie de haut niveau « plus légère » à manipuler que *DOLCE* et bien adaptée pour représenter des aspects de la « réalité sociale » (2009, p. 3). Intéressons-nous à cette ontologie prometteuse.

2.3.2.3 La découverte de DOLCE

Gangemi (2002) a conçu l'ontologie *DOLCE, Descriptive Ontology for Linguistics and Cognitive Engineering*. Nous présentons ici brièvement les caractéristiques de cette ontologie afin de justifier notre choix pour la représentation des connaissances à propos du cas *Deepwater Horizon*.

DOLCE, conçue selon la méthode *OntoClean,* a été créé au départ dans le cadre du projet *OntoWordNet*[98] pour travailler avec la base lexicale et sémantique (en anglais) *WordNet*[99]. *WordNet* est un projet toujours d'actualité mené par l'université de Princeton[100] dont l'objectif est le regroupement en jeux de « synonymes cognitifs » *(*« *synsets* ») des instances des différentes catégories grammaticales. Les liens entre ces « *synsets* » se font par le biais de relations conceptuelles d'ordre sémantique et lexicale. Ces relations permettent notamment de lever les ambiguïtés entre concepts. Voir l'exemple avec le terme « *accident* »[101].

L'objectif du projet *OntoWordNet* était principalement d'apporter une rigueur et une transparence aux catégories lexicales du haut de la hiérarchie de *WordNet*, pour les rendre exploitables dans d'autres applications, et plus généralement amener WordNet vers plus de formalisme à la manière d'une ontologie, d'être capable de vérifier la consistance de l'ensemble et de détecter d'éventuels « trous dans la raquette » (Gangemi *et al.*, 2003). En fait, c'est une ontologie qui contrôle la clarté d'une « presqu'ontologie » (une taxinomie linguistique).

L'ontologie *DOLCE* a connu de nombreuses évolutions pour arriver à la version que nous utilisons pour traiter le cas *Deepwater Horizon*. La première version présentée de *DOLCE* est quelque peu austère et peu intuitive, mais Gangemi amène les bases conceptuelles pour comprendre comment son engagement ontologique sera réalisé et ce qu'il attend de cette ontologie.

98 Source : (*Laboratory for Applied Ontology - DOLCE*, no date).
99 Source : (*About WordNet - WordNet - About WordNet*, no date; *WordNet*, no date).
100 Source : (*About WordNet - WordNet - About WordNet*, no date).
101 Source : (Exemple WordNet avec le terme accident *WordNet Search - 3.1*, no date).

Nous le citons dans le texte car c'est fondateur pour nos travaux de recherche (il s'agit du pilier O4) :

> « […] *l'objectif principal des ontologies de haut niveau est de négocier le sens des choses, soit pour permettre une coopération efficace entre de multiples agents artificiels [informatiques], soit pour établir un consensus dans une société mixte où les agents artificiels coopèrent avec des êtres humains. L'idée est de rendre aussi explicites que possible les logiques et les alternatives qui sous-tendent de tels choix, par une isolation minutieuse des options ontologiques fondamentales et de leurs relations formelles. Comme le reflète son acronyme, DOLCE a un biais cognitif clair, en ce sens qu'elle vise à saisir les catégories ontologiques qui sous-tendent le langage naturel et le sens commun humain.* » (Notre traduction, Gangemi et al., 2002, pt. 1 Introduction)

La démarche de *DOLCE* est claire : il faut trouver un accord à propos de la terminologie adoptée dans l'ontologie. Elle doit être la plus explicite possible à propos des justifications qui orientent la signification des choses du monde, cette signification étant le fruit de la négociation entre humains et non-humains. Pour cela *DOLCE* propose une rigoureuse « isolation »[102] des concepts et des relations qui permet à tous, humains et non-humains de comprendre, dès lors que la compréhension désigne « […] *un rapport qualitatif d'intellection entre une fonction mentale et les objets sur lesquels elle s'exerce* »[103], « l'étage au-dessus » de la connaissance de notre modèle *DIKU(W)*.

DOLCE cherche à avoir l'engagement ontologique le plus faible possible puisqu'elle prétend à « capturer » les fondations ontologiques du langage naturel (en matière de sémantique) et du « sens commun ». Autrement dit, il n'y aurait aucun effort cognitif à effectuer pour « rentrer » dans l'ontologie « le monde » que nous souhaitons représenter, tout semblerait « naturel » et immédiatement accessible sans un effort de conceptualisation trop important et risqué compte tenu de tout ce que nous avons déjà évoqué à propos des ontologies dans ce chapitre. De plus, afin de tenir compte des différentes analyses concurrentes d'un même cas, nous souhaitions une ontologie dont l'approche du « monde qui nous entoure » puisse nous permettre de le décrire tel que nous l'entendons et que nous retrouvons dans le caractère de ce que Gangemi appelle « *une sorte de métaphysique cognitive intrinsèquement dépendante des perceptions individuelles, des empreintes culturelles et des conventions sociales* » (Notre traduction, 2002, p. 167). Les instances issues des classes de *DOLCE* seront des représentations des « objets du monde » « sans inhérence systématique » avec cet objet. Nous pouvons donc construire toutes les représentations que nous souhaitons, mais surtout, pour un « objet du monde » donné, nous serons à même de présenter les représentations associées à cet objet proposées par d'autres que nous. C'est un point essentiel dans le choix de cette ontologie.

102 Nous le comprenons comme une rigoureuse et circonscrite axiomatisation des concepts et relations convoqués.

103 De la définition II donnée par le CNRTL (*COMPRENDRE : Définition de COMPRENDRE*, no date).

L'engagement ontologique de *DOLCE* (le pilier O_1) est décrit comme tel :

> « *« l'univers » est constitué d'événements et d'objets et ces objets participent aux événements, du moins à leur propre existence ; et sont spatialement situés.* »

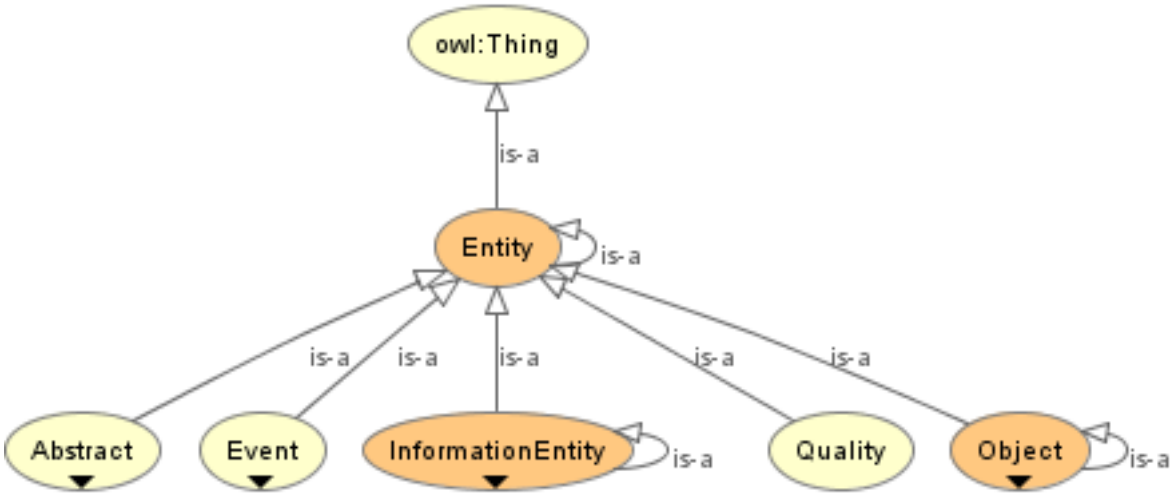
C'est ici que l'on comprend ce que l'ontologie cherche à représenter (pilier O_2) et la manière dont l'axiomatisation va se faire (pilier O_3). Regardons maintenant concrètement ce qu'il en est.

2.3.3 *DOLCE DnS UL*

Nous présentons ici la déclinaison de *DOLCE* que nous utilisons. Il s'agit de la version *DOLCE DnS UL, pour Descriptions And Situations Ultra Lite*. Par commodité, nous désignerons par *DOLCE* cette version.

Il s'agit principalement d'une simplification et d'un assouplissement de certains axiomes notamment en ce qui concerne les « régions », les espaces topologiques et les noms des classes et des relations qui ont été rendus plus accessibles. Il a été intégré le module « *Descriptions and Situations* »[104] qui permet notamment de mieux décrire les relations et la méréologie entre les descriptions et les situations dans la veine de *DOLCE*. Précisons que cette ontologie a été implémentée en langage *OWL*[105], *déclinaison* que nous allons utiliser par la suite. Nous proposons maintenant de nous intéresser aux plus hautes classes de celle-ci et de rentrer en détail dans leur description.

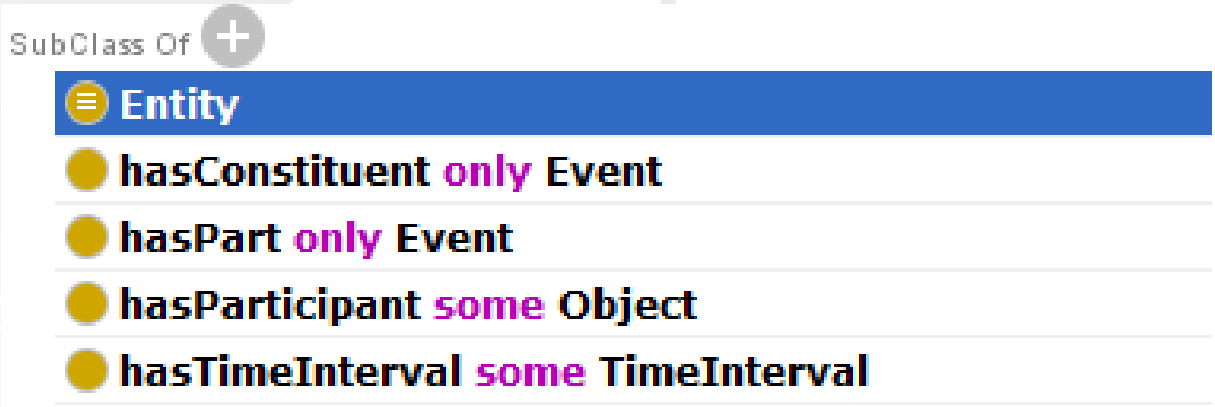
Les classes *Event*, *Object, Abstract, Quality et InformationEntity* sont les classes les plus élevées dans la hiérarchie de *DOLCE* (on ne compte pas *Entity*, qui est équivalent à l'objet d'étude et *Thing* qui est un artéfact du langage *OWL* qui n'a pas d'intérêt particulier ici). Nous les décrivons une à une et présentons leur axiomatisation, puis nous expliquerons les axiomes et les relations mises en œuvre. Voici la hiérarchie de ces classes :



104 Le fichier *xml* du module *cDnS* : http://www.ontologydesignpatterns.org/ont/cdns/cDnS.*owl*.
105 Le fichier est à l'adresse http://www.ontologydesignpatterns.org/ont/dul/DUL.*owl*.

Figure 45 : les plus hautes classes de *DOLCE DnS UL*

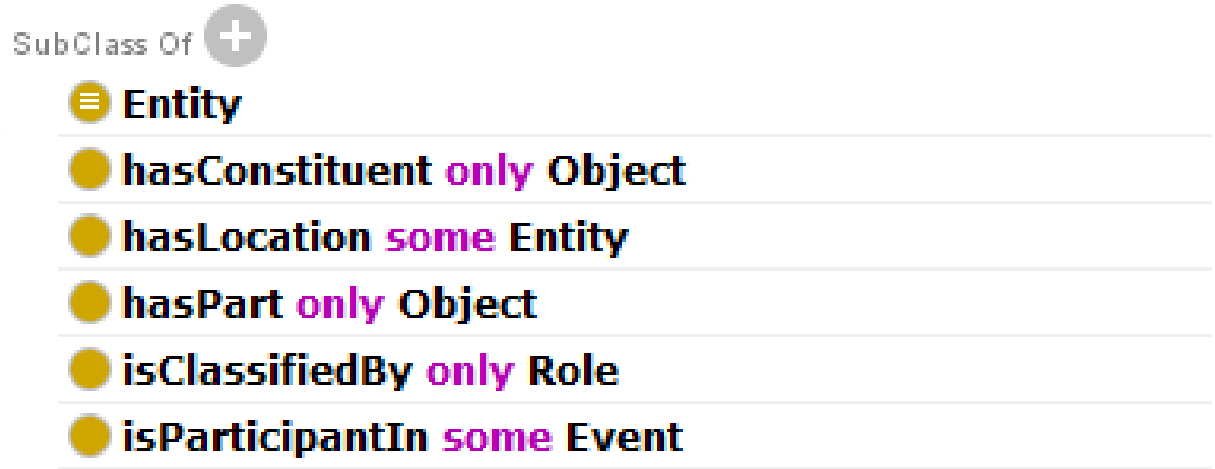
La classe *Event* est décrite par : « *Tout processus, événement ou état physique, social ou mental* »[106]. Ici, on voit que l'engagement ontologique est faible et *a priori* permet à l'utilisateur de l'ontologie de pouvoir conceptualiser l'évènement « comme il l'entend » : la suite de la description le confirme et il n'y a pas une, mais des réalités qui elles-mêmes peuvent être représentées de différentes manières parce qu'elles sont observées. De plus, *DOLCE* considère sans difficulté les différents paradigmes de réalités (classiquement naturelle et construite) et leur permet même de coexister sans



difficulté. L'axiomatisation de la classe *Event* est la suivante :

Figure 46 : les axiomes de la classe *Event*

La classe *Object* est décrite comme : « *Tout objet physique, social ou mental, ou toute substance. Suivant DOLCE, les objets participent toujours à un évènement (au moins leur propre vie), et sont spatialement situés* » Voici l'axiomatisation de la



classe :

Figure 47 : les axiomes de la classe *Object*

L'objet est le participant à l'évènement ; il peut prendre virtuellement toutes les formes possibles et les classes *Event* et *Object* sont évidemment disjointes.

106 La suite de la description est accessible dans le fichier *owl* de l'ontologie. Elle est particulièrement longue et détaillée.

La classe *Abstract* est décrite comme : « *Toute Entité qui ne peut être située dans l'espace-temps* [...] ». Cette classe contient en fait la classe *Region,* qui est la super classe des espaces topologiques. Une région est utilisée comme valeur pour la qualité d'une entité. Par exemple, *TimeInterval*, *SpaceRegion*, *PhysicalAttribute*, *Amount*, *SocialAttribute* sont toutes des sous-classes de *Region*. Elles sont des régions maximales d'un espace topologique particulier. Pour être tout à fait rigoureux, seule la classe *Event* exige *a minima* un espace topologique de la classe *TimeInterval*. La classe *Object* n'est pas axiomatiquement restreinte à la classe *Region*, mais à *Entity*, c'est-à-dire l'ensemble des classes de l'ontologie, donc aucune restriction de domaine, mais avec une exigence de localisation topologique (restriction existentielle *some*) tout de même. La classe *Abstract* n'a pas d'axiomatisation spécifique en sus de ses propriétés héritées de la classe *Entity* (c'est-à-dire ontologiquement rien sauf l'existence en tant que classe).

La classe *Quality* est décrite comme : « *Tout aspect d'une Entité, (mais pas une partie de celle-ci), qui ne peut exister sans cette Entité.* » C'est presque un clin d'œil à notre recherche tant le lien avec la définition aristotélicienne de l'accident semble figurer dans la définition de *Quality*. On comprend donc qu'il s'agit d'attributs, de qualificatifs d'autres choses qui ne sauraient exister en tant que tel. C'est le « sujet » de la classe « *Region* ». Son axiomatisation est la suivante :

Figure 48 : les axiomes de la classe *Quality*

Enfin, la classe *InformationEntity* est décrite comme : « Une information, qu'elle soit matérialisée ou non. » De même que pour la classe *Abstract*, la classe *InformationEntity* n'a pas d'axiomatisation spécifique en sus de ses propriétés héritées de la classe *Entity* (c'est-à-dire ontologiquement rien sauf l'existence en tant que classe).

A ce stade, nous pouvons proposer une représentation graphique[107] de l'engagement ontologique que nous rappelons :

> « *« l'univers » est constitué d'événements et d'objets et ces objets participent aux événements, du moins à leur propre existence ; et sont spatialement situés.* »

107 Le graphe est là encore effectué à l'aide du logiciel COE CmapTools.

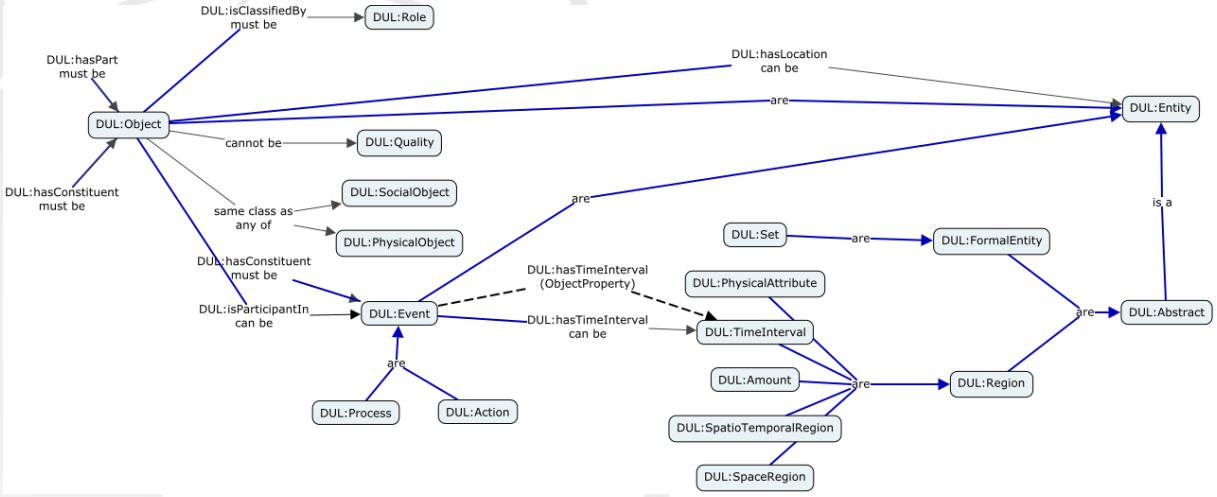


Figure 49 : l'axiomatisation de l'engagement ontologique de *DOLCE*

C'est de cette manière que nous allons décrire le cas *Deepwater Horizon*. Au-delà, nous allons voir par la suite qu'il s'agit de la structure d'analyse qui va être notre référentiel axiomatique pour notre apport de thèse[108].

*

* *

Dans ce chapitre 2, nous avons exposé en quoi le domaine scientifique de l'ingénierie des connaissances est celui dans lequel nous devions nous aventurer pour chercher une réponse au problème soulevé à la suite de notre chapitre 1. Nous avons découvert le concept d'ontologie et exposé en profondeur la conceptualisation, les critères de conception, la construction des axiomes et la logique des prédicats comme support pour une formalisation de ce concept philosophique et de son atterrissage « pas si en douceur que cela » dans la communauté scientifique au vu des nombreuses et encore actuelles discussions et controverses, moins à propos de leurs utilités et utilisations que sur leurs natures même et leurs portées. Nous avons constaté que les ontologies à l'heure actuelle sont un support conceptuel majeur aux développements de l'ingénierie des connaissances notamment par le biais du Web sémantique et des données liées. Nous avons présenté quelques exemples de réalisation à ce propos et des résultats préliminaires concernant notre traitement du cas *Deepwater Horizon*. En continuant notre travail sur les ontologies et en nous inscrivant dans la pensée de Guarino, nous avons fait le choix d'une ontologie de haut niveau, *DOLCE*, créé par Gangemi, pour apporter une solution à notre problème. Nous avons montré en quoi le choix de *DOLCE* est pertinent et exposé les capacités de « modélisation » de cette ontologie.

Nous montrerons en détail dans les chapitres suivants l'étendue de « l'instanciation » du cas *Deepwater Horizon* à l'aide de *DOLCE*. La représentation ontologique est en soi un résultat de recherche, cependant, c'est aussi au cours de cette

108 Notamment pour la production de graphes de causalité.

appropriation de l'ontologie que sont nés deux axes de recherches importants de notre travail de thèse :

- le premier est l'étude d'une piste pour créer de la connaissance depuis un document vers une ontologie, en interrogeant la pertinence du sens à donner à la représentation du cas que nous étudions ;
- le second, qui découle du premier et du constat des limites ontologiques et de *DOLCE* à propos de l'expression de la causalité, est une étude des pistes algorithmiques pour détecter automatiquement la causalité dans un texte et de la transcrire dans une ontologie.

Le chapitre 3 a pour objectif de présenter nos réflexions sur ces deux thèmes.

Chapitre 3 Algorithmes de population automatique d'une ontologie d'accident

Nous positionnons notre recherche dans le domaine du Traitement Automatique du Langage Naturel (TALN) ou *Natural Language Processsing* (*NLP*) et particulièrement dans ce qu'il est de plus en plus convenu d'appeler *Natural Language Understanding* (*NLU*). Nous mobilisons des concepts hérités de la linguistique (particulièrement la linguistique formelle et la grammaire structurelle), de l'ingénierie des connaissances et des ontologies dont nous avons fait la présentation dans le chapitre 2, et des avancées en matière d'algorithmie rendues possibles ces dernières années par le développement des algorithmes d'apprentissage. Les recherches effectuées notamment par le laboratoire *NLP* de l'université de Stanford[109] ont été notre point de départ dans l'élaboration des travaux qui vont être présentés.

Dans ce chapitre, nous allons présenter notre proposition de recherche : une suite d'algorithmes pour la population automatique des ontologies, en particulier les ontologies d'évènements qui traitent de l'accident. Nous allons d'abord présenter l'élaboration d'un algorithme *Named Entity Recognizer* dont l'objectif est de trouver des instances d'évènements dans un texte écrit (3.1). Nous présentons en premier les fondamentaux de linguistique morphosyntaxique et fonctionnelle, base de notre algorithme (3.1.1). Puis nous élargissons notre approche du texte par la sémantique et le processus de lemmatisation (3.1.2). Enfin, nous montrons la conception de l'algorithme *NER*, les résultats attendus et les limites théoriques (3.1.3). Pendant de l'événementialité de l'accident, nous allons ensuite nous pencher sur l'étude de la causalité et de son traitement algorithmique (3.2). Nous présentons d'abord l'aspect théorique de la causalité et de son lien à l'évènement (3.2.1). Ensuite, nous faisons un exposé de l'existant en matière de traitement du langage naturel par une Intelligence Artificielle et rappelons les fondamentaux sémantiques, mais aussi le travail humain en amont de l'algorithmie (3.2.2). Nous finissons cette sous-section par la présentation de notre approche de la causalité exprimée par une méthode bayésienne de forage de causalité que nous avons mise au point (3.2.3). Enfin, nous faisons la présentation de notre machine, apte à répondre à la question pourquoi ? et lier causalité et événementialité (3.3). Nous abordons le concept de cheminement causal, clé de la compréhension du texte et résultat souhaité (3.3.1). Puis nous faisons la démonstration théorique de la faisabilité de notre machine (3.3.2) et enfin, nous montrons les capacités attendues de cette machine, ayant conscience de l'expressivité très particulière du raisonnement causal contrefactuel, par une projection d'utilisations possibles (3.3.3).

109 Le site du laboratoire (*The Stanford Natural Language Processing Group*, no date).

Ce chapitre est une première proposition théorique et méthodologique avec une visée clairement opérationnelle pour explorer une solution en deux temps :

- cerner des expressions d'évènements dans un texte et créer une ontologie de ces évènements avec leur chronologie ;
- cerner la causalité exprimée dans un texte et être capable de délivrer un graphe de causalité situé dans la chronologie évènementielle.

L'objectif est, qu'à terme, l'utilisateur puisse interroger le texte étudié pour obtenir des réponses pertinentes à deux questions : quand et pourquoi un évènement est survenu ? En sus de lui faire gagner beaucoup de temps en lui économisant la lecture (ou en la ciblant mieux), il pourra exécuter ce logiciel sur différents textes et confronter les résultats. Nous espérons qu'il sera possible pour l'utilisateur de « forer » la causalité à un degré de profondeur supérieur à une lecture « traditionnelle » et, tout du moins, de pouvoir rattacher la causalité exprimée à l'événementialité relatée.

La méthode que nous avons développée s'articule comme suit :

- utilisation d'une ontologie de haut niveau pour servir de référence axiomatique pour la recherche morphosyntaxique et fonctionnelle d'expressions susceptibles de décrire des évènements, construction d'un algorithme *NER* dans cet objectif ;
- détection sémantique et syntaxique de phrases exprimant la causalité par le biais d'algorithmes bayésiens pour un traitement déterministe des arguments de causalité ;
- structuration d'une réponse pertinente à un questionnement causal par l'utilisation d'algorithmes de similarité sémantique et de surface ;
- liaison entre causalité et événementialité dans une représentation graphique par la mise en œuvre d'un processus itératif de questionnements.

A travers les réflexions sur la sémantique du premier chapitre, nous avons constaté que pour rapprocher des concepts (au sens linguistique), il est efficace, d'une part, de réduire les dimensions sémantiques des unités lexicales considérées (réduction de l'espace sémantique) et, d'autre part, de simplifier le « domaine de définition », c'est-à-dire de considérer ou non l'ensemble des éléments qui le peuple. Le domaine de définition est ici circonscrit à un texte écrit ; l'unité d'analyse est la phrase, séquence de mots bornée par deux ponctuations fortes. La phrase est donc elle-même composée d'un ensemble ordonné de morphèmes, que nous appellerons « mots » dans la suite de ce chapitre.

Nous appelons « paradigme de bienveillance » le fait que nous supposons vrai que la grammaire est juste, que l'orthographe est respectée et que toute phrase a un sens ou une explication à son sens dans le domaine de définition. Ainsi, nous

supposons que le principe de compositionnalité est respecté[110] et nous supposons aussi que l'hypothèse de sémantique distribuée est vraie[111].

Nous supposons également qu'il n'y a pas de catastrophe de sens à l'échelle du morphème. Évidemment, nous avons bien conscience du caractère polysémique de la langue et nous supposons que l'utilisation d'un lemme, quelles que soient les multiples morphologies adoptées dans le texte, ne change pas, sauf explication. La première signification amenée par la volonté de l'auteur lors de la première utilisation dans le texte reste identique et ce, quel que soit le renvoi lexical du lemme en question. Travaillant particulièrement sur des données issues d'enquêtes post-accident (produites dans des rapports, des minutes de procès, des articles de science, …), nous pouvons supposer que les significations portées par ces lemmes resteront identiques en « tout point » du support desdites données (dans le domaine de définition). Par exemple, un *cofferdam*, sans autre qualification textuelle (sémantique ou relationnelle), à partir du moment où il a été défini tel que l'objet technique qui permet de capturer un effluent sous-marin (nous allons le voir par la suite), nous considèrerons par la suite que cette même unité lexicale pointera toujours vers la même signification. Quand bien même il sera qualifié par la suite comme « le *cofferdam* de BP » ou le *cofferdam* descendu dans le fond de l'eau » ou « le *cofferdam* est peint en blanc », l'unité *cofferdam* est maintenue dans son sens premier donné par l'auteur et la co-textualité pourra amener (induit) des variations sans pour autant changer le noyau de sens[112].

Dans ce chapitre, sauf explicitement mentionné, nous instancierons notre domaine de définition dans un texte particulier, le *staff working paper n°6*[113] écrit par le personnel de la commission nationale chargée du rapport d'enquête sur l'accident de *Deepwater Horizon* et dont le destinataire est le Président des États-Unis[114]. Ce document écrit (il y en a au moins six) est l'un des préliminaires au rapport officiel[115] et contient donc des informations très intéressantes pour notre réflexion sur le cas. Chaque *staff working paper* est axé sur une thématique particulière de l'enquête post-accident ; le n°6 est intitulé : *Stopping The Spill : The Five-Month Effort To Kill The Macondo Well*. Ce texte fait trente-neuf pages et expose en profondeur cette ingénierie de l'extrême mise en œuvre pour tenter de tuer le puits *Macondo* et stopper la fuite d'hydrocarbures. Par la suite, nous l'utiliserons en totalité ou plus spécifiquement en nous concentrant notamment sur sa partie *B Cofferdam*, relatant l'histoire du

110 Nous allons y revenir.
111 L'hypothèse de la sémantique distribuée consiste à poser comme vrai que « *des mots avec des significations similaires apparaîtront avec des [mots] voisins similaires s'il y a suffisamment de texte disponible* » (Schütze and Pedersen, 1995). Il y a des associations de mots (co-occurrences) plus fréquentes que d'autres qui composent les expressions et les phrases.
112 Voir le chapitre 1 à ce propos.
113 Source : (National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, 2011b).
114 Les noms des membres de la commission sont donnés dans le rapport (National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, 2011a, p. iv).
115 Source : (National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, 2011a).

cofferdam. Cet extrait comporte trente-huit phrases et est mis en annexe également. Le numéro de phrase correspond à l'ordre d'apparition dans cet extrait en particulier.

3.1 Un algorithme de population automatique des ontologies d'accident

Nous présentons ici ce qui pose peut-être le plus de problèmes à l'heure actuelle dans le domaine de l'ingénierie des connaissances : la population des ontologies. Le schéma suivant illustre ce que nous entendons par « peupler » une ontologie :

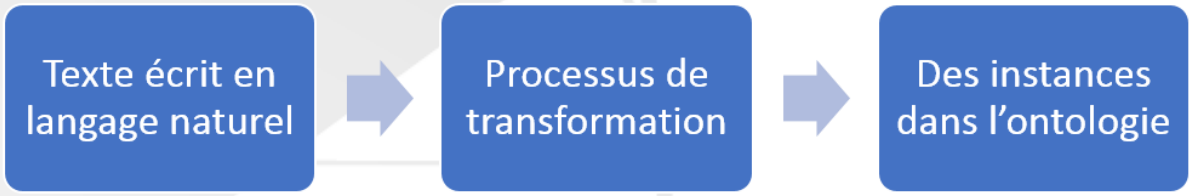


Figure 50 : la population des ontologies

Nous travaillons avec l'ontologie *DOLCE*, particulièrement bien adaptée à notre matériau de recherche, le texte écrit[116]. Aussi, nous considérons le processus de population de l'ontologie comme un processus de transformation depuis le langage naturel vers le langage ontologique (ou graphe ontologique). Avant cela, nous pouvons d'ores et déjà affirmer, par expérience, que la population d'une ontologie est une tâche difficile qui fait l'objet d'une recherche récente, mais encore éparse, en particulier dans le champ des sciences de l'information, de l'ingénierie des connaissances ou encore de l'informatique. Citons les travaux de Alani et al. (2003), Amardeilh (2006), Davalcu et al. (2003), Geleijnse and Korst (2005), Jupp et al. (Jupp *et al.*, 2011, 2012), Venkataraman and Mendonca (2003), Wolstencroft et al. (2013b, 2013a, 2012, 2011) qui tentent tous d'apporter une méthode (parfois concrétisée dans un outil informatique) pour faciliter ou automatiser cette tâche. La biologie et ses disciplines associées (bio-informatique, génétique) sont les plus en demande pour ce type de recherches et il y a un vide béant concernant les autres disciplines. Passons ensuite sur les quelques outils développés pour aider à produire des instances en masse tels que *Populous* de Jupp et al. (2011) ou encore *Ontopop* de Amardeilh (2006) pour la création d'annotations. En fait, ces outils ne permettent pas de détecter dans le texte des entités quelconques pouvant faire de bons candidats pour des instances de classes ; nous sommes encore réduits à peupler l'ontologie à la main ; seule la production d'instances déjà définies par l'humain est accélérée par une diminution de la répétition, et encore, pour des entités représentant des instances de classes « simples » à évaluer (la classe *NaturalPerson* par exemple dans *DOLCE*). Parmi les travaux les plus prometteurs dans la phase d'extraction et d'automatisation de données écrites, citons ceux d'Alani et al. (2003), avec le projet *ArteQuakt*, et ceux de Davalcu et al. (2003), avec le logiciel *OntoMiner*. Il semble toutefois que le projet *ArteQuakt* soit resté à l'état de projet et nous n'avons pas réussi à tester et évaluer le logiciel *OntoMiner*, introuvable en ligne malgré nos nombreuses recherches[117].

Pourtant, le constat est unanime, la population « à la main » d'une ontologie est un processus à la fois fastidieux et risqué. Fastidieux face à la masse de données à

116 Voir chapitre précédent, à la section *DOLCE*.
117 Il y a pourtant quelques papiers, Draghici (2003) et Khatri (2004, 2007) et un site web (Davalcu *et al.*, 2003) mais nous n'avons jamais réussi à avoir accès au logiciel.

considérer et risqué compte tenu des innombrables erreurs possibles, que ce soit d'ordre méthodologique, ontologique ou épistémologique.

Pour avancer, nous proposons d'aborder la question, non pas en essayant de comprendre « cognitivement » l'ontologie pour s'approprier l'engagement ontologique, mais plutôt de partir de l'ontologie pour explorer une ressource textuelle. Nous proposons, en effet, de déterminer un processus de mise en correspondance de l'expression formelle d'une classe de l'ontologie avec une structure textuelle, de l'ontologie vers le texte et retour. Notre approche revient à *construire un pont entre la structure de l'ontologie et la structure syntaxique de la source*.

La réalisation de ce pont est l'élaboration d'un algorithme dit *Named Entity Recognizer* (*NER*), qui, à partir d'une phrase, sera capable de déduire des instances et des relations possibles en rapport avec des classes et des prédicats d'une ontologie et *in fine* de proposer des candidats à la population de l'ontologie.

Ce choix est justifié par le fait que les documents auxquels nous nous intéressons sont principalement des analyses structurées relatives à des accidents industriels (rapports d'investigations, articles scientifiques ou techniques) ; ils visent à éclairer un fait, expliquer une séquence ou justifier une recommandation. Ils sont donc de nature argumentative. Nous formons ainsi l'hypothèse que la « connaissance » capturée par la structure de l'ontologie correspond *a priori* à la forme des arguments saillants de ces argumentations. Nous développons en outre cette approche dans une perspective d'apprentissage statistique ; compte tenu du nombre limité de structures argumentatives constatées à la lecture des rapports d'investigations techniques, il est raisonnable d'espérer que l'ensemble des structures mises en exergue dans la base de connaissance progressivement enrichie converge vers l'ensemble des structures d'argumentation utilisées dans les communications techniques.

Pour mémoire, l'axiomatisation de l'engagement ontologique de *DOLCE*[118] se présente ainsi :

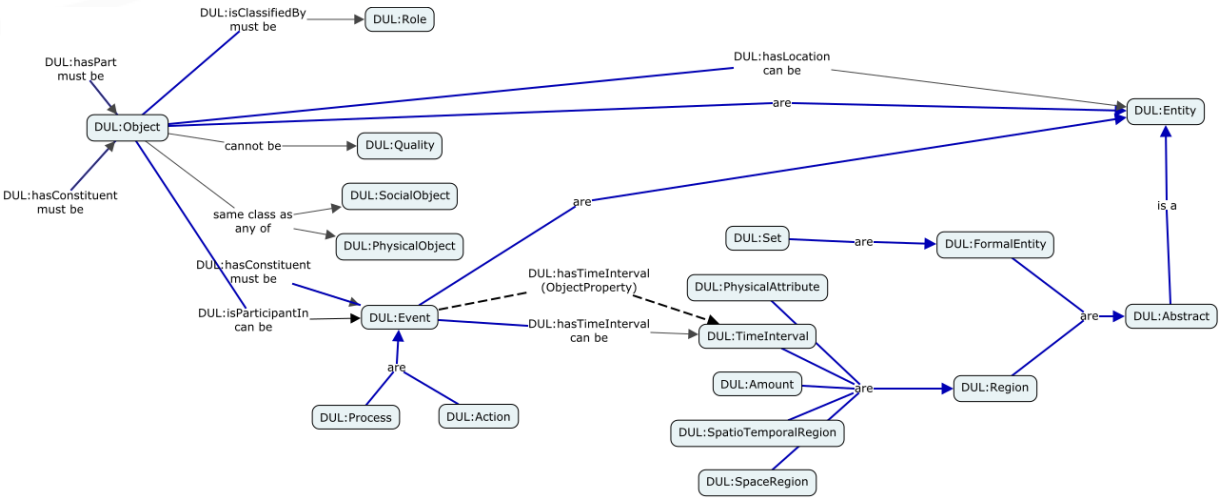


Figure 51 : l'axiomatisation de l'engagement ontologique de *DOLCE*

118 Voir le chapitre 2.

Afin de faciliter la démonstration, nous réduisons « le spectre » à la classe *Event* qui est, de toute façon, la classe fondamentale de notre processus et nous avons :

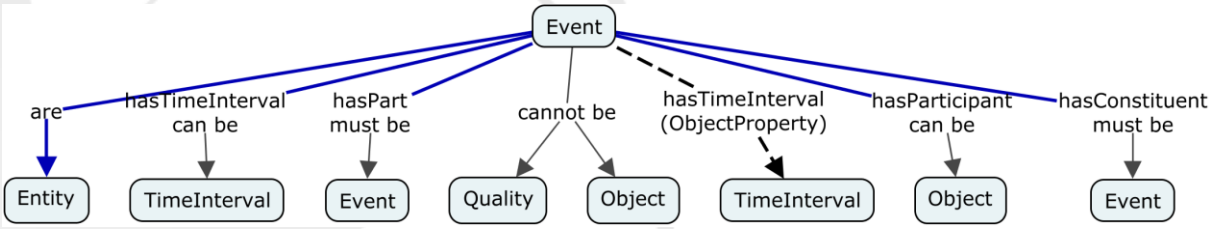


Figure 52 : l'axiomatique de la classe *Event* dans *DOLCE*

Si l'on réduit encore l'axiomatique aux axiomes existentiels (ceux qui sont générateurs de la classe), qui possèdent une restriction existentielle exprimée par *can be, some* ou encore ∃ (« il existe »), on a la classe *Event* telle que :

- Event hasTimeInterval *some* TimeInterval
- Event hasParticipant *some* Object

Ce sont les deux axiomes fondamentaux dans la définition de la classe *Event*.

De la même manière, nous devons nous intéresser aux axiomes fondamentaux qui définissent les classes *TimeInterval* et *Object*. Pour la classe *TimeInterval*, il n'existe pas d'axiome existentiel en rapport avec la classe *Event*.

Et pour la classe *Object* :

- hasLocation *some* Entity ;
- isParticipantIn *some* Event (fonction inverse de hasParticipant).

On obtient donc la description axiomatique existentielle de la classe *Event* qui est :

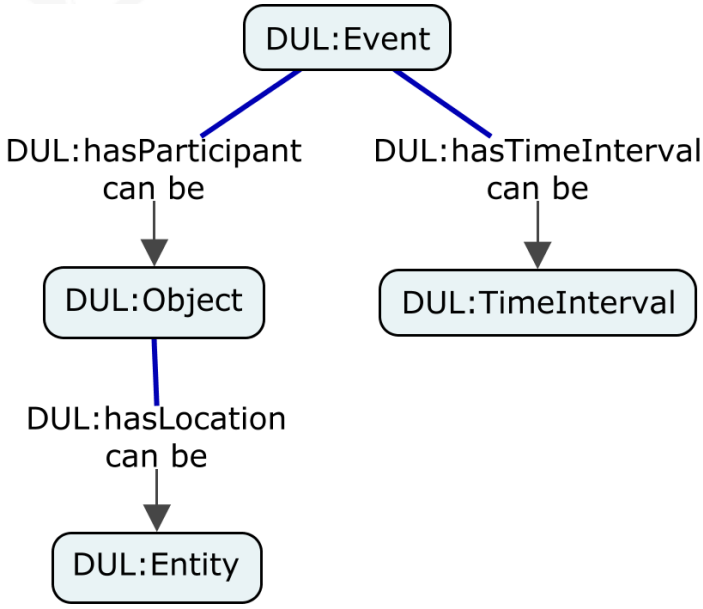


Figure 53 : la description axiomatique existentielle de la classe Event

Un évènement ontologique est donc constitué d'une instance de la classe *Event* qui comprend des instances des classes liées par l'axiomatique. Un évènement diffère d'un autre lorsque ses classes diffèrent.

Nous avons défini une structure ontologique existentielle d'une classe. Notre objectif est de réaliser un pont entre cette structure et celle des phrases d'un texte. Nous

décrivons d'abord la structure morphosyntaxique et fonctionnelle sur laquelle nous nous appuierons.

3.1.1 Structure morphosyntaxique et fonctionnelle

Précisons dès à présent que c'est bien notre regard sur l'axiomatique d'une ontologie qui nous a donné l'idée de réfléchir à la recherche d'une structure dans un texte. Cette structure sera la syntaxe de la phrase. Ce parti-pris nous permet de nous affranchir[119] d'une analyse sémantique extrêmement coûteuse en ressources (informatiques, cognitives) avec toutes les considérations de l'approche « purement sémantique » que cela nécessiterait de traiter. Nous voulons un algorithme rustique, parcimonieux, bref reposant sur des règles heuristiques qu'un marin ne renierait pas. Rien n'empêchera par la suite de lui adjoindre des « guides sémantiques » pour cibler au laser la recherche de cette structure ou de l'enrichir avec les très rares moteurs sémantiques d'extraction d'information comme *FRED*[120] (Gangemi *et al.*, 2013, 2016) qui sont eux aussi encore au stade du développement.

Pour revenir à l'analyse morphosyntaxique et fonctionnelle, nous acceptons sans discuter les concepts linguistiques que nous sollicitons. Nous nous intéressons à la branche informatique de la linguistique, elle-même héritière de la logique de Richard Montague, qui « *propose en effet un système syntaxico-sémantique complet et authentiquement compositionnel, constitué d'un ensemble fini de règles syntaxiques, chacune associée à une règle sémantique* » (Tellier, no date, p. 75).

L'analyse morphosyntaxique et fonctionnelle a pour objectif de décomposer une phrase en catégories grammaticales puis de déterminer les fonctions de ces catégories dans la phrase. Le principe élémentaire d'analyse est celui de la substituabilité (que l'on avait déjà rencontré dans le chapitre 1 pour la définition de la synonymie) qui consiste à considérer que deux unités sont grammaticalement équivalentes si et seulement si le remplacement de l'une par l'autre ne change pas le critère de grammaticalité. Intéressons-nous maintenant à notre matériau de recherche, la phrase. Nous précisons que nous introduirons au fil de l'eau les implémentations informatiques existantes qui convoquent les principes linguistiques et permettent d'opérer automatiquement les différentes procédés mis en œuvre dans l'analyse morphosyntaxique et fonctionnelle.

3.1.1.1 La phrase, unité de l'analyse

Le principe de compositionnalité stipule que « *le sens d'une proposition ne dépend que du sens des mots qui la constituent et de leur mode de combinaison syntaxique* » (Tellier, no date, p. 74 et suivantes). Tellier parle de proposition à propos de ce que l'on pourrait considérer naïvement comme phrase. On considèrera pour la suite qu'une proposition est un constituant logique de la phrase (au sens de la logique des prédicats et dans la réflexion à propos de l'émergence de la représentation logique

119 Pas strictement car il y aura besoin tout de même d'un référentiel sémantique seulement exploité d'un point de vue morphologique.

120 Nous avons eu l'occasion de pouvoir l'essayer avec des résultats tout à fait intéressants, mais il semble que le moteur soit indisponible depuis quelques mois à l'adresse : http://wit.istc.cnr.it/stlab-tools/fred.

de la connaissance qui passe par la vérification des conditions pour qu'une proposition soit vraie « dans le monde ») et donc qu'il peut y avoir plusieurs propositions dans une phrase et que les relations entre ces propositions sont des relations d'ordre logique. En fait, il y a différentes écoles de pensée pour la définition de la phrase, mais nous inscrivant plutôt dans une approche structuraliste du langage écrit, nous choisissons la phrase comme un « *Assemblage de mots caractérisé par la complétude sémantique, la cohésion grammaticale et par les pauses qui l'entourent (à l'écrit par les signes de ponctuation forte)* »[121]

Et plus particulièrement comme un « *Assemblage de mots caractérisé par son autonomie grammaticale* »[122].

La phrase, quelle que soit sa proposition en termes de logique, est autonome dans sa signification et peut donc nous servir de support, nécessaire et suffisant, pour la mise en place de notre parallèle théorique avec l'axiomatique de l'ontologie. Voyons maintenant le traitement de la phrase au travers du principe de compositionnalité. Commençons par l'approche sémantique.

3.1.2 Sémantique : lemme, lexique et lemmatisation

Le lemme est une « forme » d'un mot et précisément la forme « de base », celle qui sert de référence, celle du lexique et cette forme est choisie, désignée ; elle fait consensus : « *Un même mot se manifeste dans les langues flexionnelles sous différentes formes. Celles-ci portent des marques morphologiques qui sont définies par les lois flexionnelles de la langue en question. L'ensemble de ces formes constitue le paradigme de flexion ou paradigme flexionnel du mot, et la forme choisie par convention dans cet ensemble est le lemme du mot. Le lemme est donc une forme canonique choisie dans le paradigme qui représente l'ensemble des mots qui sont liés entre eux par la flexion* » (Cailliau, 2010, p. 32).

Ici, nous considérons le « mot » comme une forme quelconque du lemme, c'est donc un morphème[123] du lemme en question.

La composition introduit de nouveaux lemmes à partir d'autres déjà disponibles. Elle consiste à mettre bout à bout des morphèmes pour en créer de nouveaux :

$$porte + conteneurs \rightarrow porteconteneurs$$

Le morphème nouvellement formé possède son propre sens même si dans ce cas, intuitivement, on comprend qu'il hérite aussi de la sémantique de ses composants. Cependant, on ne pourrait expliquer l'entièreté du sens du morphème « porte-conteneurs » avec seulement les deux morphèmes lexicaux qui le composent ; en effet, dans le cas présent, le porte-conteneurs est un type de navire, ce qui ne peut

121 Source : (Définition : C. – LINGUISTIQUE 1. [Unité ling. définie au moyen de critères variés] a) GRAMM. CLASS « PHRASE : Définition de PHRASE, » n.d.).

122 Source : (Définition : b) GRAMM. STRUCTURALE ou FONCTIONNALISTE « PHRASE : Définition de PHRASE », n.d.)

123 « Un morphème est une unité linguistique minimale ayant une forme et un sens » (Tellier, no date, p. 23).

se voir *a priori* dans les deux morphèmes qui le composent. « Porte-conteneurs » doit être considéré comme un lemme nouveau. La composition génère des lemmes nouveaux par transformation morphologique, mais ne garantit en rien la nouvelle sémantique comme héritière des anciennes.

L'affixation, quant à elle, est l'ajout de morphèmes grammaticaux à des morphèmes lexicaux. Il y a deux types d'affixation, dérivationnelle et flexionnelle. La dérivation est l'ajout d'un affixe (préfixe ou suffixe) à un « mot » dont l'objectif est pour le coup de changer le sens du « mot ». On crée donc un nouveau morphème. Mais l'approche est plutôt lexicologique que morphologique dans le sens où la dérivation crée un « mot » nouveau :

$$sous + marin \rightarrow sousmarin$$

La flexion, par ailleurs, est la variation morphologique d'un « mot » autour du temps, de la personne, du genre et du nombre sans que cela ne provoque un changement de sens du mot en question. La conjugaison des verbes est typiquement productrice de flexions. Contrairement à la composition et à la dérivation, la flexion ne change pas la sémantique du morphème « de départ ». Le lemme reste le même.

Dans la suite, nous nous appuierons sur des lexiques, ou recueil de lemmes. Le plus connu et utilisé à l'heure actuelle est *WordNet*[124] que nous avons présenté au chapitre 2.

3.1.2.1 Les algorithmes de lemmatisation

A l'heure actuelle, il existe des algorithmes de lemmatisation. Ils effectuent le processus de lemmatisation et proposent pour chaque « mot » d'une phrase une étiquette probable d'appartenance à un groupe grammatical. Deux des plus utilisés sont *TreeTagger*[125] et *Stanford Log-linear Part-Of-Speech Tagger*[126].

Nous avons déjà utilisé l'algorithme de l'université de Stanford sur les données que nous avons créées pour représenter la connaissance scientifique du cas *Deepwater Horizon*[127]. Nous recherchons, à partir du lemme « *accident* » (en anglais), toutes les formes utilisées dans ce texte compilé. Nous avons les résultats dans le tableau suivant :

Word	POS	Count	Word	POS	Count
accident	Noun	447	**accident-prone**	JJ	1
accident	NN	228	**accident/_/disaster**	Noun	1
accidents	NNS	209	**accident/_/disaster**	NN	1
Accidents	NNS	10	**Accidental**	Proper Noun	1
Accident	Proper Noun	40	**Accidental**	NNP	1

124 Source : (*WordNet*, no date).
125 Sources : (*TreeTagger - a language independent part-of-speech tagger | Institute for Natural Language Processing | University of Stuttgart*, no date; *TreeTagger*, no date; Schmid, 1994, 1995).
126 Sources (Toutanova and Manning, 2000; Toutanova *et al.*, 2003).
127 Voir le chapitre 1.

Accident	NNP	40	**accidentally**	Adv	1
accidental	Adj	35	**accidentally**	RB	1
accidental	JJ	30	**Accidents**	Proper Noun	1
Accidental	JJ	5	**Accidents**	NNP	1
post-accident	Adj	2	**fire-accident**	Adj	1
post-accident	JJ	2	**fire-accident**	JJ	1
accident-prone	Adj	1			

Tableau 15 : les formes issues du lemme « accident » dans un texte compilé d'abstracts d'articles de science portant sur le cas *Deepwater Horizon*

On trouve donc tous les formes du lemme *accident* présents dans le texte ainsi que le nombre d'occurrences. A chaque forme est associée sa catégorie grammaticale[128] et quelques précisions sur d'éventuelles flexions (singulier ou pluriel typiquement) :

accidental	**Adj**	**35**

On a ici *accidental*, une forme fléchie qui a comme catégorie grammaticale « adjectif » du lemme *accident* et elle apparaît à 35 reprises dans le texte[129].

Ou encore :

post-accident	**JJ**	**2**

Post-accident est une forme dérivée avec un préfixe « post- » qui change le sens pour donner au « mot » « post-accident » l'idée de « qui suit temporellement l'accident ».

Purnelle (1996) précise notamment que la lemmatisation permet de supprimer les formes homographes des lemmes dont l'étymologie ou l'emploi sont différents : c'est un processus de levée d'ambiguïtés sémantiques très puissant. Concernant le processus de lemmatisation, on pourra lire également Abeillé et al. (2003)[130] qui explique en français comment s'effectue l'annotation morphosyntaxique dans la langue. Nous ne nous intéressons donc pas à l'extension sémantique du « mot », mais seulement à son « *intension* » indispensable au processus de lemmatisation.

3.1.2.2 Les relations grammaticales et la grammaire universelle

Comme nous l'avons présenté auparavant, la syntaxe est soumise à une grammaire dont les règles arrangent l'ordre des « mots » dans une phrase en rapport à leur nature et à leur emploi. Ces règles déterminent, en sus de la sémantique propre à chaque mot et toujours selon le principe de compositionnalité, le sens de la phrase.

Chaque langue possède sa grammaire. Cependant, dans le domaine du TALN, des travaux ont été menés dont l'hypothèse de recherche est l'existence de relations de grammaire qui pourraient avoir un caractère universel, c'est-à-dire que l'on retrouverait dans l'immense majorité des langues (l'approche linguistique croisée). Il

128 Nous allons voir par la suite comment la catégorie grammaticale est déterminée.
129 Il y a également quelques erreurs, souvent dues à la présence d'une lettre en majuscule en début de mot, que l'algorithme peut parfois interpréter comme un nom propre plutôt que la mise en majuscule de la première lettre d'une unité lexicale lorsque celle-ci débute une phrase.
130 Nous allons présenter par la suite en profondeur ces travaux.

a donc été question de réflexions pour développer ce qu'on appelle une grammaire hors contexte, plus souvent citée comme *Free-Context Grammar* (*FCG*). L'une des grammaires hors-contexte les plus utilisées est celle des *Universal Dependencies*[131] (*UD*). Cette grammaire a été retravaillée par la suite et une extension, ou plutôt un raffinage, a été proposé par le biais des *Stanford Dependencies* (*SD*) et *Stanford Enhanced Dependencies* (*SD Enhanced*) ; l'ensemble des travaux est présenté sur le site du laboratoire *NLP* de l'université de Stanford[132] notamment dans les papiers de De Marneffe et al. (2014) et Manning et al. (2014) à propos de l'évolution des *SD* vers les *SD Enhanced*. De Marneffe et Manning (2008) présentent et expliquent également toutes les relations grammaticales existantes ainsi que leur hiérarchie, les cas particuliers ou ambigus. Encore une fois, le travail du laboratoire *NLP* de l'université de Stanford est en très bonne adéquation avec nos besoins et leurs codes sont disponibles, expliqués et tenus à jour.

La nomenclature adoptée à l'heure actuelle est donc celle des *UD* enrichies des *SD Enhanced*. Le site[133] présente un ensemble statistique tenu à jour pour chaque étiquette grammaticale et chaque relation. Le répertoire de référence *Universal Dependencies*[134] présente une matrice des relations en fonction des articulations qu'elles amènent et de leurs dépendants ainsi qu'une liste explicative de toutes ces relations.

En 2018, il y a trente-sept relations universelles :

	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	list parataxis	orphan goeswith reparandum	punct root dep

Tableau 16 : matrice de référence des relations grammaticales universelles

131 Source : (*UD_English*, no date). L'algorithme PASSAGE de l'INRIA présenté succinctement auparavant est, par exemple, une autre manière d'analyser les fonctions grammaticales dans une phrase.

132 Source : https://nlp.stanford.edu/

133 Source : http://universaldependencies.org/treebanks/en_ewt/index.html

134 Source : http://universaldependencies.org/u/dep/all.html

Voici la relation *nsubj* explicitée par différents exemples :

Figure 54 : la relation *nsubj*

Elle marque les sujets nominaux d'une clause. Les sujets sont directement dépendants du prédicat principal de la clause, qui peut être un verbe, un nom ou un adjectif » (« nsubj, » n.d.).

La compréhension de ces relations nécessite un certain temps car il ne faut pas oublier qu'elles ont été conçues dans une visée universelle et donc qu'intuitivement, il y en a quelques-unes qui peuvent paraître loin du mode de pensée du lecteur. Certaines semblent même être quelque peu « artificielles », justement dans le but de « lisser » des relations qui sont similaires, mais pas rigoureusement identiques dans différentes langues (comme la relation *case*).

3.1.2.3 Parsing d'une phrase

Nous présentons maintenant l'implémentation informatique qui permet l'analyse morphosyntaxique et fonctionnelle : ce sont des algorithmes appelés *parsers*[135]. Il faut bien se rendre compte ici que ce type d'analyse sur le langage naturel

135 Par hypothèse pour la suite, le *parser* reconnaît la phrase comme unité syntaxique complète dans un corpus.

est très poussé et que le travail sur ces algorithmes est l'un des fers de lance de la recherche actuelle en TALN. Nous devons citer les travaux de l'université de Berkeley[136] ainsi que ceux proposés par l'université Paris Diderot[137] et par l'INRIA, mais nous utiliserons essentiellement les algorithmes proposés[138] par l'université de Stanford[139].

Le *parser*, fait appel au *POS tagger* qui lui-même se sert d'une « base de données » d'apprentissage indispensable pour pouvoir identifier et proposer des structures aux phrases qu'on lui soumet (c'est un modèle statistique qui a évolué vers un modèle à réseaux de neurones[140]). La base de données étiquetées la plus utilisée est *Penn Tree Bank*[141] *;* mais pour le français, il existe une base de données étiquetées appelé *French Tree Bank*, produit des travaux d'Abeillé et al. (2003; 2004). L'INRIA, propose son *parser FRMG* (FRench Meta-Grammar) pour l'analyse du français développé par De la Clergerie et al. (2008) construit à l'aide de l'algorithme PASSAGE (projet PASSAGE[142]). En utilisant le *parser* FRMG sur notre phrase « modèle » on obtient :

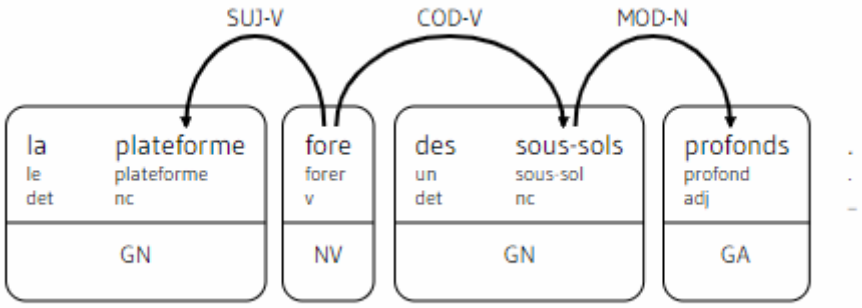


Figure 55 : analyse de l'algorithme « Passage » sur une phrase simple

Nous faisons apparaître ici les concepts sémantiques et syntaxiques que nous avons évoqués auparavant. On a donc :

- « La plateforme », nom commun, groupe nominal (syntagme nominal) sujet du verbe « fore » ;
- « fore », forme du lemme « forer », noyau verbal de la phrase ;
- « des sous-sols », groupe nominal complément d'objet direct du verbe « fore » et composé de « des », forme du lemme « un », déterminant et « sous-sols », nom commun ;

136 Source : (*The Berkeley NLP Group*, no date).
137 Source : (*Accueil | Corpus arboré pour le français*, no date).
138 Source : (*L'analyseur syntaxique FRMG pour le français | FRMG Wiki*, no date).
139 Source : (*The Stanford Natural Language Processing Group*, no date) les problèmes des quelques autres *parsers* que nous avons essayés sont principalement l'ergonomie ou la difficulté d'accès à l'algorithmie ou au *tagging*. En l'occurrence, le *tagging* initial dans *Penn Tree Bank* est effectué grâce à *TreeTagger*.
140 Source : (Chen and Manning, 2014).
141 Sources : (*Penn Treebank II Tags*, no date; *Penn Treebank P.O.S. Tags*, no date; Santorini, 1990)
142 Source : (*Passage: ANR MDCA Passage*, no date). La réflexion informatique à la création de cet algorithme dépasse largement le cadre de nos compétences.

- « profonds », adjectif, groupe adjectival « modifieur[143] » du groupe nominal « des sous-sols ».

Nous avons présenté la structure syntaxique d'une phrase et avons montré quels sont les méthodes et moyens pour effectuer une analyse morphosyntaxique et fonctionnelle et leurs implémentations informatiques à l'heure actuelle par des algorithmes que sont les *parsers*. Nous venons également de montrer que certains *parsers* sont très puissants et arrivent à déterminer la structure grammaticale d'une phrase, ses composants et leurs fonctions. Nous avons vu également en quoi l'utilisation d'une grammaire « universelle » permet de s'affranchir assez bien du langage dans lequel est écrite la phrase étudiée. Cela nous permet donc de faire émerger une structure syntaxique « universelle » et ses propriétés. Nous souhaitons maintenant mettre en parallèle ces propriétés avec l'axiomatique de l'ontologie.

3.1.3 Un algorithme NER (Named Entity Recognizer)

Un algorithme *NER* (*Named Entity Recognizer*) est un algorithme qui traite des séquences de mots et est capable d'identifier des entités et de les classer avec un étiquetage sémantique. Par exemple, la séquence « Thad Allen » pourrait être reconnue comme « Personne », ou « Houston » comme « Ville » ou « Lieu ». Il s'agit donc de déterminer une « classe »[144] à laquelle appartient l'entité analysée.

Le laboratoire *NLP* de l'université de Stanford propose un *NER* apte à reconnaître quelques classes (Personne, Organisation, Date ou Lieu) de façon assez robuste (Finkel, Grenager and Manning, 2005) :

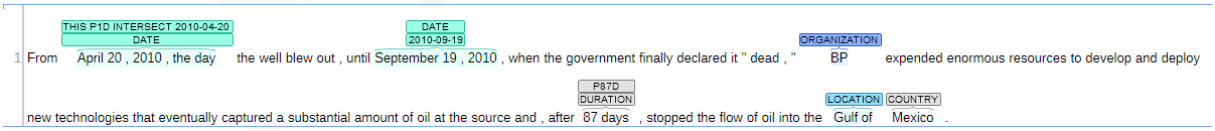


Figure 56 : résultat de l'algorithme *NER* de l'université de Stanford sur une phrase.

Les entités détectées ont été correctement identifiées.

Cet algorithme n'est pas conçu de la même manière que celui que nous allons proposer, mais nous pensons que la partie qui traite des expressions de temporalité devra être considérée pour la suite de nos travaux. Nous allons exploiter les concepts et outils d'analyse de surface (*surface pattern analysis*) et d'analyse relationnelle (*dependencies pattern analysis*) pour construire une structure syntaxique cible. Ce sont les expressions utilisées en TALN et cela correspond à l'analyse morphosyntaxique, d'une part, et à l'analyse fonctionnelle des fonctions grammaticales dans la phrase, d'autre part.

Soit la phrase suivante, que nous considérons comme un modèle – au sens de candidat type – pour alimenter une instance de la classe *Event* dans *DOLCE* :

« *On May 10, 2010, the Junk Shot manifold was lowered to the sea floor*. »[145]

143 Barbarisme utilisé dans les papiers du projet PASSAGE.
144 Exactement dans la même veine qu'une classe ontologique, voir le chapitre 2.
145 Extrait du compte rendu de la phase deux du procès MDL2179, (Judge Barbier, 2015).

En effet, si l'on revient à l'axiomatique de la classe *Event*, on a l'intuition que cette phrase « rentre bien » dans les critères de la classe relativement à un certain objet. Intuition confirmée par le recoupement avec la lecture de la description en langage naturel de la classe *Object* que nous rappelons :

> « *[…] les objets participent toujours à un évènement (au moins leur propre vie) et sont spatialement situés.* »[146]

La phrase étudiée remplit les conditions axiomatiques imposées par la classe « *Event* » : on a bien ici un « objet », « *the Junk Shot manifold* »[147], qui vit une existence (il est le sujet d'un verbe) décrite au moyen du verbe employé « *was lowered* » qui implique explicitement une temporalité portée par la conjugaison du verbe (l'action est terminée). Cette temporalité est également portée par un complément circonstanciel de temps, « *On May 10* », qui fixe cet évènement sur une chronologie calendaire. Enfin, l'espace topologique est ici physique et il est décrit à la fois implicitement par le verbe et explicitement par le complément circonstanciel de lieu « *the sea floor* ».

Il va s'agir maintenant de le démontrer par l'analyse morphosyntaxique et fonctionnelle et d'être capable par la suite de déterminer, si elle existe, une structure syntaxique spécifique à cette classe de l'ontologie. Si l'on revient à l'algorithme *NER* de l'université de Stanford, on a le résultat suivant :

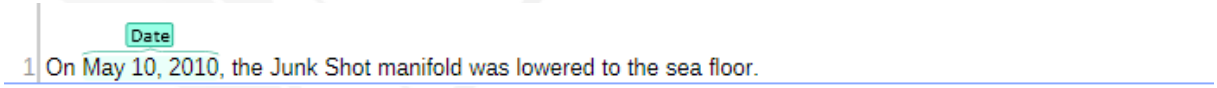


Figure 57 : résultat de l'algorithme *NER* l'université de Stanford

L'algorithme reconnaît l'entité *May 10, 2010*, comme une date, ce qui est juste. Cependant, malgré la simplicité de la phrase, il ne détecte aucune autre entité susceptible d'exister dans le texte.

Nous allons donc proposer un algorithme *NER* dont la capacité sera la détection de « phrases » candidates à des instances de la description axiomatique existentielle de la classe *Event*. Cet algorithme sera construit sur la base d'un *parser* dont les résultats seront analysés *via* la structure ontologique que nous allons définir.

3.1.3.1 Recherche d'une structure syntaxique cible

Si nous revenons à notre phrase modèle – « *On May 10, 2010, the Junk Shot manifold was lowered to the sea floor* » –, en utilisant la suite algorithmique de l'université de Stanford *NLP Core*[148], on a les résultats de l'étiquetage grammatical :

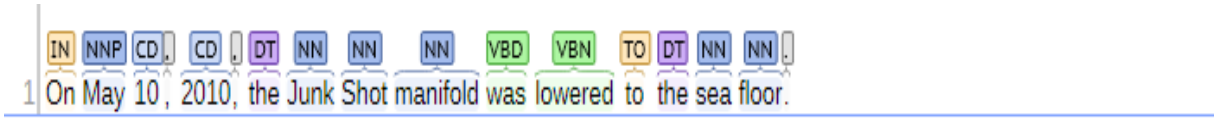


Figure 58 : résultat de l'algortihme *POS tagger* de l'université de Stanford

146 Source : http://www.ontologydesignpatterns.org/ont/dul/DUL.*owl*.

147 Un équipement nécessaire pour la mise en œuvre d'une procédure spécifique pour tenter de tuer le puits.

148 Source : (*Stanford CoreNLP – Natural language software | Stanford CoreNLP*, no date).

On retrouve la nature des éléments de la phrase. Il est intéressant de noter que le *POS tagger* considère le noyau verbal en deux éléments conjugués au passé (étiquettes *VBD* et *VBN*) ; on a une information au-delà de l'élément qui est le temps de la phrase et cette information est donnée par la forme du noyau verbal.

Regardons maintenant l'arbre syntagmatique de la phrase « modèle » :

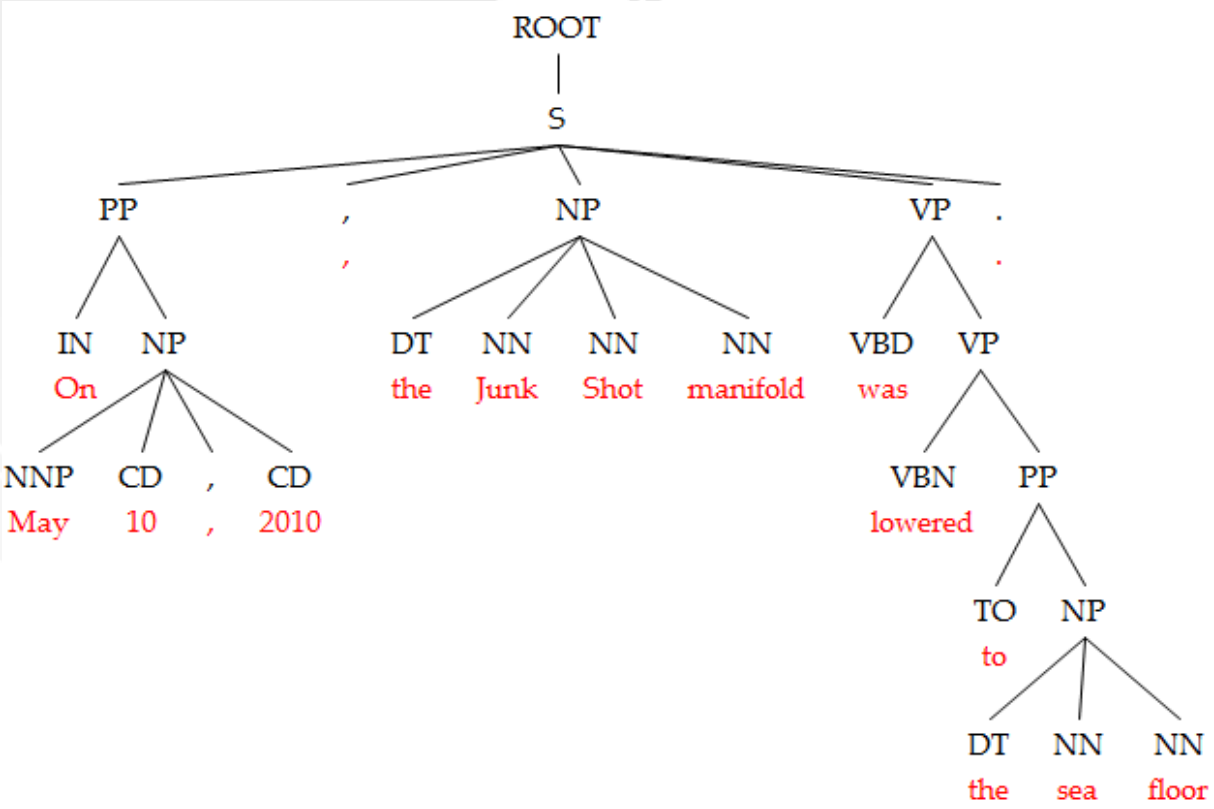


Figure 59 : arbre syntagmatique obtenu par le *parser* de l'université de Stanford

Nous avons ici quatre syntagmes[149] : un prépositionnel, un syntagme nominal et un noyau verbal. Ce dernier étant composé d'un verbe conjugué et d'un autre noyau verbal lui-même composé d'un verbe au participe passé et d'un syntagme prépositionnel (le quatrième syntagme). L'arbre syntagmatique permet de se rendre compte de la profondeur (l'enchâssement en grammaire) d'une phrase. On voit plus facilement qu'il y a trois « boîtes » (dont une plus « profonde » que les autres) reliées les unes aux autres et que celles-ci ont une fonction à assurer dans la phrase.

Regardons maintenant l'analyse fonctionnelle (effectuée selon la grammaire universelle *UD Enhanced*) :

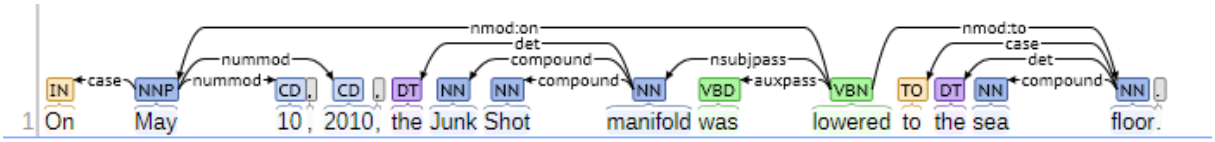


Figure 60 : résultats du *parser enhanced SD* de l'université de Stanford

149 « Un syntagme est [...] un groupe de mots qui correspond à un sous-arbre d'un arbre d'analyse syntaxique complet. » (Tellier, no date, p. 38) ou encore « une combinaison de morphèmes ou de mots qui se suivent et produisent un sens acceptable », particulièrement « un groupe d'unités linguistiques significatives formant une unité dans une organisation hiérarchisée de la phrase. » , définition B (*SYNTAGME : Définition de SYNTAGME*, no date).

On a les relations :

- *case* : on a deux relations de ce type qui introduisent des modificateurs nominaux qui jouent un rôle par rapport au noyau verbal et plus précisément par rapport au verbe conjugué. Deux cas sont introduits ici, un cas introduit par *on* et un par *to* ;
- *nummod : numeric modifier* : un « modifieur » numérique d'un nom est une expression numérique qui sert à modifier le sens du nom avec une quantité. Ici, l'entité *May* est modifiée deux fois, par *10* et par *2010*. On commence à y voir comme une structure de date ;
- *nmod : on* : *nominal modifier (enhanced with : on) :* la relation *nmod* est utilisée pour les dépendants nominaux d'un autre nom ou d'une autre expression de nom et correspond fonctionnellement à un attribut ou à un complément génitif. Dans ce cas, il est précisé qu'il y a une préposition et qu'il s'agit de *on* ;
- *nmod : to* : *nominal modifier (enhanced with : to) :* identique à la définition au-dessus , mais le « modifieur » est introduit par la préposition *to* ;
- *det* : il s'agit ici de la relation entre la tête d'un nominal et son déterminant : ici *the* avec *manifold* ;
- *compound* : la relation composée est l'une des trois relations pour les expressions à multiples mots (*MultiWords Expressions, MWE*). Ici, par exemple, on a *Junk* et *Shot* qui sont reliés à *manifold* pour former le syntagme nominal *Junk Shot manifold* ;
- *nsubjpass* : un sujet nominal (*nsubj*) est le sujet syntaxique de la phrase. La précision *pass* (grâce aux relations *SD Enhanced*) correspond à la voix passive employée. Ici, *the Junk Shot manifold* est le sujet à la voix passive du verbe *lower* ;
- *auxpass* : un auxiliaire *aux* d'une phrase est un mot de fonction associé à un attribut verbal qui exprime des catégories telles que le temps, l'humeur, l'aspect, la voix ou l'évidentialité[150]. Il est précisé que l'auxiliaire est employé à la voix passive. Ici, avec le temps et la voix, on comprend que l'action qui se déroule est révolue et circonscrite dans l'espace considéré (la relation *nmod : on* « enferme » l'action du verbe dans la date du 10 mai 2010).

Nous avons maintenant la structure grammaticale complète de notre phrase « modèle » (natures et fonctions des morphèmes et des syntagmes). Nous avons même, grâce à l'algorithme *NER* de l'université de Stanford, une piste sérieuse à propos du groupe de mots « *May 10, 2010*, » (nous ne parlons pas ici de syntagme car l'algorithme *NER* n'étiquette pas la préposition *on*) identifié comme une entité de *Date*.

150 Pour une explication de ces termes de grammaire, voir le papier de Barbet et Saussure (2012).

3.1.3.2 Le pont théorique entre ontologie et syntaxe

A partir de cela, nous souhaitons proposer le pont théorique entre l'analyse morphosyntaxique et fonctionnelle et l'axiomatique existentielle de la classe *Event* :

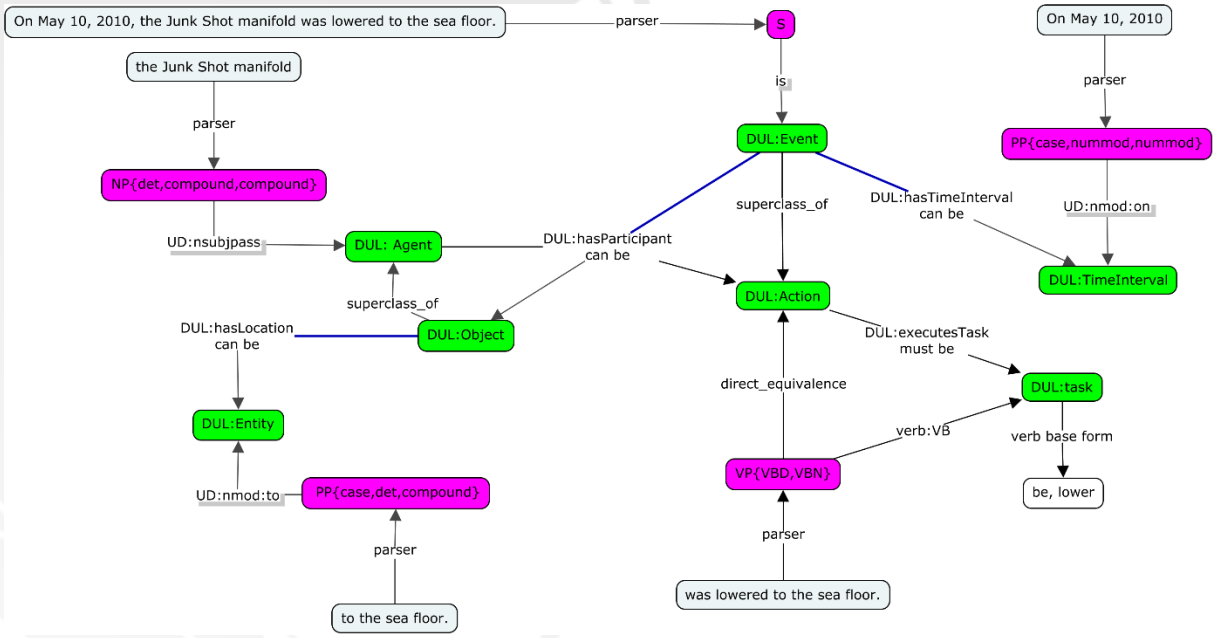


Figure 61 : pont théorique entre la syntaxe et l'ontologie *DOLCE*

L'élément syntaxique de référence sera le syntagme. Nous avons donc associé chaque structure syntaxique extrait du *parsing* avec une classe de l'axiomatique *Event* de l'ontologie *DOLCE* par rapport à la fonction grammaticale que le composant principal du syntagme prend dans la phrase complète en rapport avec le noyau verbal de cette phrase (syntagme *VP* le plus élevé dans le *parsing ET* contenant les formes verbales). Nous précisons que nous choisissons la classe *Agent* pour la fonction *UD : nsubj,* qui est une sous-classe de la classe *Object* (c'est donc ontologiquement correct) et qui porte dans sa définition la notion d'agentivité qui sied bien à la fonction sujet dans une phrase.

De la même manière, on créera des instances dans les classes *Action* et *Task*. *Action* hérite des propriétés de *Event* (sous-classe), en ce sens, on considèrera le noyau verbal complet comme l'action en question, dans ce cas « *was lowered to the sea floor* ». La classe *Action* précise aussi la nature de l'objet participant (*Agent*) et ajoute le concept de *Task* que nous transcrirons ici comme les formes infinitives (*VB : Verb Base form*) des formes verbales présentes dans le noyau verbal.

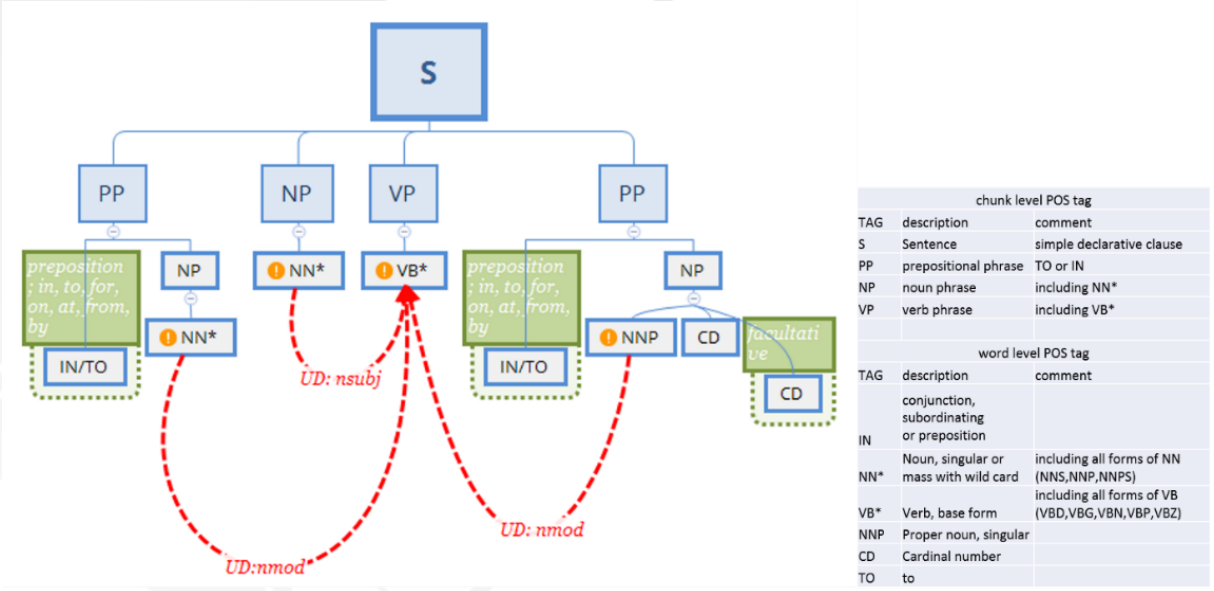
Il est donc théoriquement possible d'associer des séquences de mots avec des classes de l'ontologie et donc de considérer ces séquences de mots comme des instances des classes en question. Et au-delà, on le voit sur l'illustration précédente, nous pouvons faire correspondre les fonctions grammaticales et, plus généralement, la syntaxe à l'axiomatique de la classe.

On pourra donc définir comme une instance de classe une phrase (séquence complète) si et seulement si toutes les instances de classes trouvent un candidat dans les séquences du *parsing*.

Nous cherchons maintenant à proposer une structure syntaxique spécifique à l'axiomatique de la classe *Event*, une sorte de *pattern* structurel « universel ».

3.1.3.3 Etablir une structure syntaxique cible de la classe Event

Nous allons proposer maintenant une structure syntaxique cible à détecter dans un document écrit par le biais de notre algorithme *NER*. Voici la structure syntaxique cible que nous proposons :



chunk level POS tag		
TAG	description	comment
S	Sentence	simple declarative clause
PP	prepositional phrase	TO or IN
NP	noun phrase	including NN*
VP	verb phrase	including VB*

word level POS tag		
TAG	description	comment
IN	conjunction, subordinating or preposition	
NN*	Noun, singular or mass with wild card	including all forms of NN (NNS,NNP,NNPS)
VB*	Verb, base form	including all forms of VB (VBD,VBG,VBN,VBP,VBZ)
NNP	Proper noun, singular	
CD	Cardinal number	
TO	to	

Figure 62 : la structure *DOLCE Event* selon la grammaire universelle *UD*

Cette structure est donc celle que va rechercher l'algorithme dans un texte écrit pour tenter de déterminer des instances de classe *Event* de *DOLCE*. Il y a deux couches d'analyse proposées.

En premier lieu, la couche purement syntaxique. La structure syntaxique cible est composée d'un syntagme nominal *NP* dont la fonction est sujet par la relation universelle (*UD : nsubj*) d'un noyau verbal *VP* dont le forme verbale *VB** n'a pas d'importance. Pour les syntagmes prépositionnels, le minimum à rechercher est d'un par composition : un syntagme *IN/TO, NP* (qui jouera la fonction de complément circonstanciel de lieu) et un syntagme *IN/TO, NN*, CD, (CD)* (qui jouera la fonction de complément circonstanciel de temps). La fonction de chaque syntagme sera donc complément circonstanciel de lieu ou de temps du noyau verbal avec la relation universelle (*UD : nmod*). Le composant *IN/TO* recouvre les adpositions, c'est-à-dire les prépositions et postpositions. Ces « mots-outils » (*cover term*) introduisent ou terminent un complément qui leur est subordonné[151] et que l'on retrouve dans la relation universelle *UD : case*. En français, par exemple, il n'y a pas de postposition.

En second lieu, nous étudions la couche syntaxico-sémantique (signalée par le point d'exclamation dans le schéma). En effet, si l'on souhaite affiner un peu plus la recherche, il faut considérer la langue utilisée et nous pouvons proposer des « guides sémantiques » qui vont probablement améliorer la détection de la structure syntaxique cible pour la classe *Event*. Cela va concerner surtout la manière d'identifier les

151 Source : (*Adposition*, 2015).

syntagmes prépositionnels qui jouent les fonctions de compléments circonstanciels. Voici des guides sémantiques qui nous semblent pertinents pour le moment. Pour le syntagme [*IN/TO, NN*, CD, (CD)]* : en anglais, les mois de l'année ont une majuscule, alors on peut proposer le composant *NNP* (nom propre au singulier) tel que [*IN/TO, NNP, CD, (CD)]*. De plus, on pourra considérer que le deuxième composant *CD* (*cardinal number*) pourra être omis puisqu'une date peut s'exprimer telle que *May 1985* (on considère l'année), ou *May 03* (on considère le jour). Néanmoins, une date correctement exprimée sera bien de la forme *NNP, CD, CD* telle que *May 03, 1985*. Pour les syntagmes *PP*, on pourra effectuer une recherche doublement guidée en fonction des prépositions. Le premier guide est justement de limiter la recherche des adpositions uniquement aux prépositions et le second sera de cibler certaines prépositions, beaucoup plus susceptibles d'introduire des compléments de temps ou de lieu, telles que : *in, to, for, on, at, from* et *by*. La mise en place de cette deuxième couche d'analyse sera proposée comme une ouverture à cette thèse tant le travail est déjà conséquent pour le déploiement de la première.

3.1.3.4 Le Csas, référentiel pour l'algorithme

Nous appelons *parallel checking* la capacité de notre algorithme *NER* à comparer le résultat du *parsing* d'une phrase (*POS tagging*, arbre syntagmatique, relations fonctionnelles *UD* enhanced) avec un référentiel préétabli pour être capable de déterminer si la phrase (ou la proposition) est un candidat à l'instance de classe. La structure syntaxique correspondant à l'axiomatique d'une classe de l'ontologie est donc ce référentiel que nous allons réécrire comme un ensemble de conditions appelé *Csas* soit *Class syntactic-axiomatic structure*. Nous proposons que la vérification suive la règle simple de la nécessité et de la suffisance (la parcimonie de la ressource). Par exemple, et pour revenir au besoin du processus, voici le tableau des conditions à vérifier dans une phrase (ou une proposition) pour que celle-ci puisse prétendre à être une instance de la classe *Event* dans *DOLCE* :

Event Csas
S[2(PP), NP, VP]
VB*/*UD* : nmod/NN*
VB*/*UD* : nsubj/NN*
VB*/*UD* : nmod/NNP
PP[ADP, NNP, CD, CD]

Tableau 17 : les conditions *Csas* à vérifier pour une instance de classe *Event* dans *DOLCE*

Le *workflow* de l'algorithme *NER* est présenté en page suivante.

Le *workflow* de l'algorithme

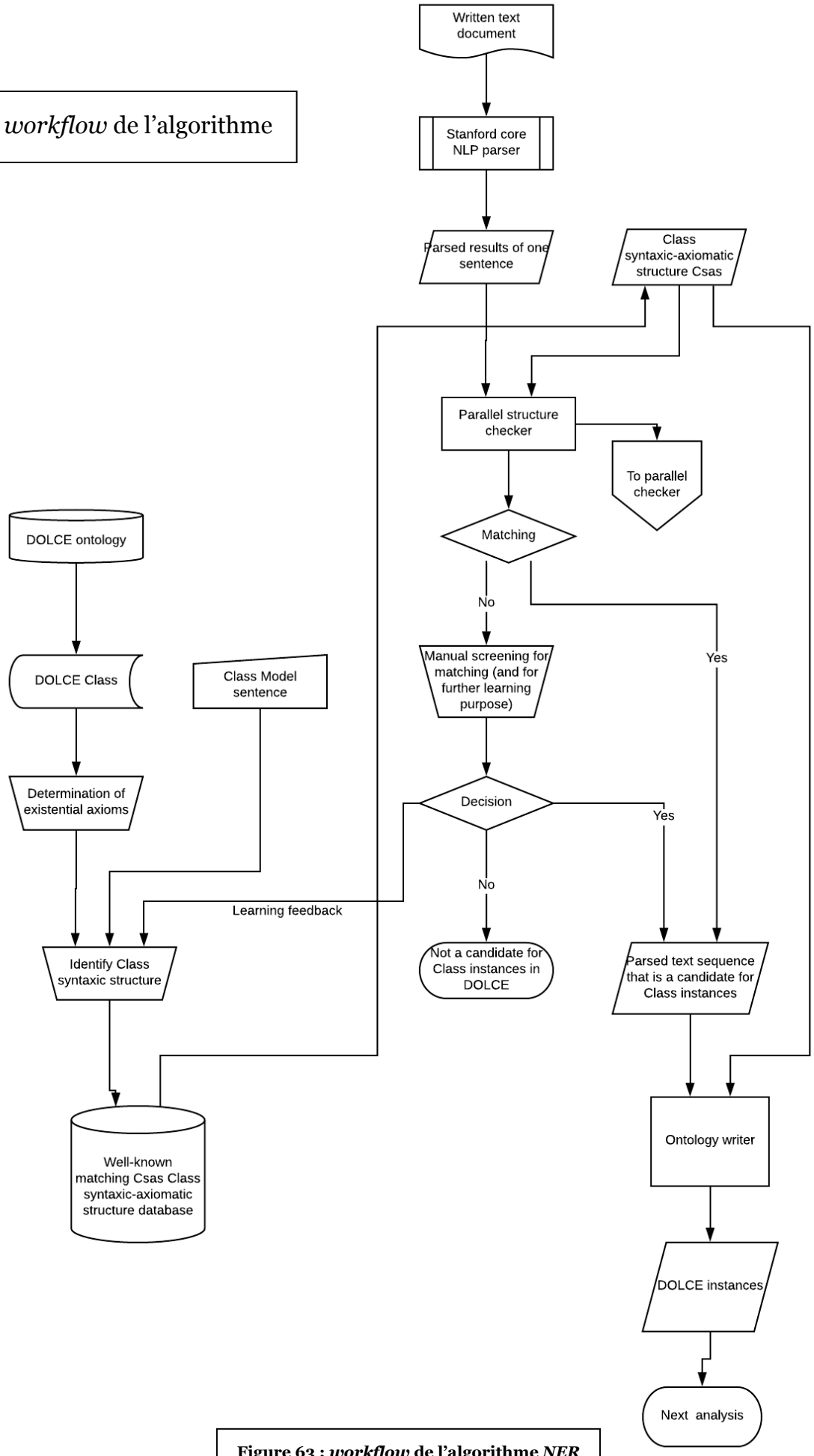


Figure 63 : *workflow* de l'algorithme *NER*

Et une vision du processus *parallel checking* effectué par l'algorithme :

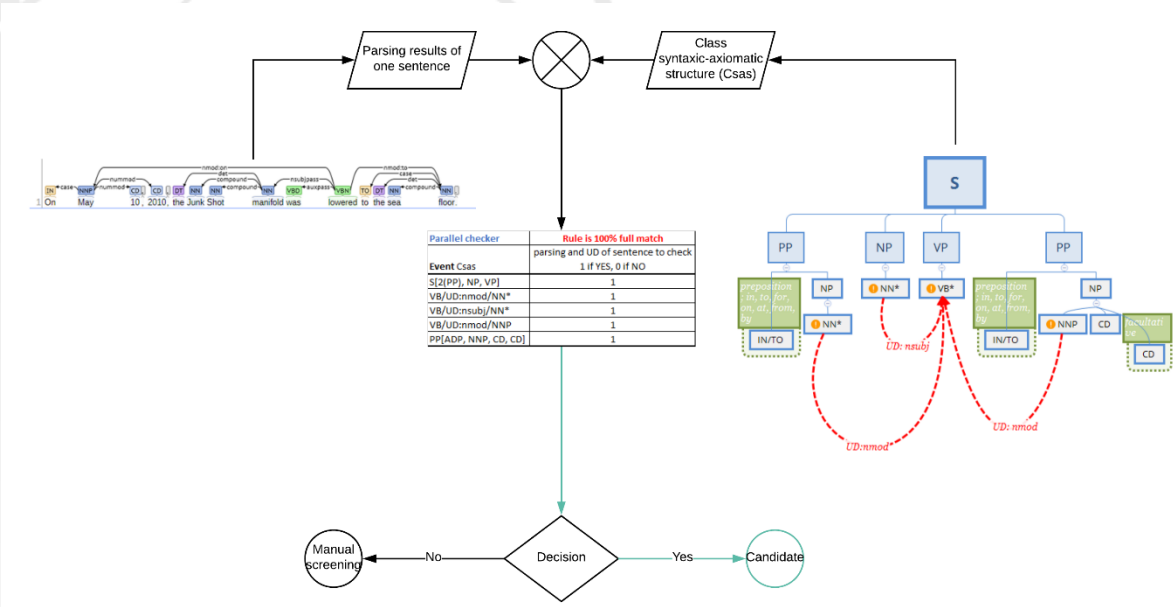


Figure 64 : fonctionnement du *Parallel checker*

Le *parallel checker* agit comme un comparateur entre la structure syntaxique ciblée (le *Csas* sélectionné) et la structure syntaxique identifiée dans la phrase à analyser. En fonction des éléments identifiés, il accorde un score, dont on pourra déterminer les règles par la suite, qui permet de classer la phrase comme une phrase portant une structure identique au *Csas* que l'algorithme vérifie et donc, *in fine*, de classer la phrase comme une phrase d'évènement et d'en extraire les instances associées.

Ontologiquement, suivant *DOLCE*, on ne pourra considérer comme étant une instance candidate de la classe *Event* une phrase dont la structure syntaxique ne répond pas *a minima* à l'ensemble des conditions *Csas*. Ne pas le faire reviendrait à remettre en question l'ontologie de l'évènement amenée par *DOLCE* ce qui n'est pas l'objectif de cette thèse, bien au contraire, puisque nous la considérons comme un référentiel « indiscutable ».

3.1.3.5 Variations et difficultés potentielles de mise en œuvre

Etant donné que nous ne proposons qu'une version théorique de notre algorithme *NER*, nous étudions ici quelques difficultés potentielles. Nous explicitons également le traitement particulier de l'ordonnancement chronologique des évènements dans l'ontologie.

Les résultats attendus seront de la forme :

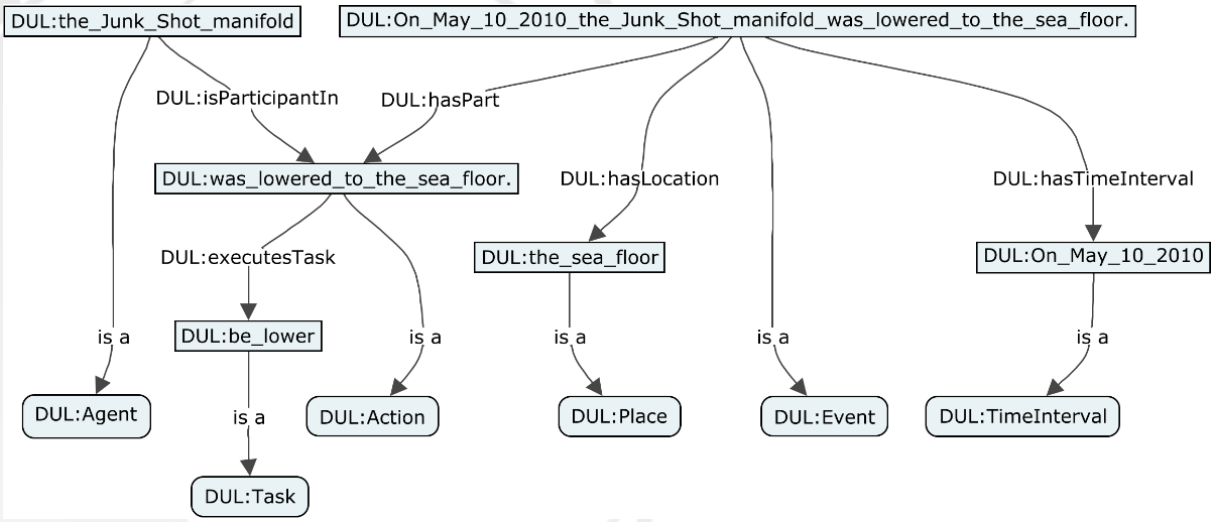


Figure 65 : vue d'artiste des instances écrites dans l'ontologie *DOLCE* par l'algorithme *NER*

L'ensemble des instances créées dans les différentes classes, en commençant par l'axiomatique de la classe *Event,* forme la population de l'ontologie. Nous avons donc mis au point une méthode algorithmique, implémentée par le biais d'un algorithme *NER*, qui permet de peupler une ontologie à partir d'un texte écrit. Nous allons voir maintenant les éventuels problèmes auxquels nous pensons être confrontés et nous présenterons l'exploitation des instances d'évènement pour la réalisation d'un graphe d'évènement.

3.1.3.5.1 Structure minimale acceptable

Mentionnons d'abord qu'il est possible que la structure syntaxique des phrases à analyser soit plus conséquente que le *Csas* recherché. Nous proposons pour le moment d'extraire seulement les éléments de la structure syntaxique cible. Dans le cas où il y aurait plusieurs fois la structure syntaxique cible, cas d'une phrase complexe ou avec un degré d'enchâssement élevé, on extraira exactement les séquences détectées. L'importance est sur l'exactitude de la structure identifiée dans son ontologie. L'ontologie n'est pas précisément une structure tolérante puisque, par définition, les instances créées suivent exclusivement les principes de subsomption (héritage) et de disjonction (appartenance).

Par ailleurs, l'algorithme *NER* va travailler sur chaque phrase du texte et il est donc possible qu'il extraie des séquences de mots sensiblement identiques et dont il faut s'assurer que c'est effectivement le cas pour éviter la création de doublons dans les instances ou au contraire, la fusion de séquences de mots en une instance alors qu'elles ne sont pas identiques. Ce cas est plus difficile à traiter. Nous allons évaluer les séquences candidates aux instances *Agent*, *Place*[152]*, TimeInterval, Action* et *Event* pour déceler d'éventuels doublons. Cette comparaison est fondamentale car elle permet de fusionner toutes les séquences candidates en une instance de classe unique parce qu'elles sont ontologiquement identiques.

152 Sous-classe de la classe *Entity* qui correspond à la relation *hasLocation*.

Nous recourrons à la similarité de surface, qui permet d'identifier des séquences identiques en fonction de la granularité d'analyse (caractère, mot, phrase) et de donner un score d'évaluation en fonction de la méthode utilisée[153] (« What is text similarity ? | Kavita Ganesan, » n.d.). La similarité s'exprime de deux manières :

- on mesure une similarité absolue entre deux unités, « jusqu'où sont-elles similaires ? » (On évalue le degré de similarité entre deux unités) ;
- on mesure la proximité relative entre *n* unités, « quelles sont les unités les plus proches entre elles ? » (On effectue un classement, un *ranking*).

De nombreux algorithmes ont été créés pour effectuer ces deux types de comparaisons en fonction des besoins[154], notamment décrits par H. Gomaa and A. Fahmy (2013) et Vogler (2013).

Le site du Pr. Bill Buchanan[155] propose (entre autres) de nombreux outils de mesure de similarité de texte en ligne utilisé dans le domaine *computer forensics*. Nous avons choisi d'utiliser l'algorithme de Smith–Waterman (1981) car « *Il est utile pour les séquences dissemblables dont on soupçonne qu'elles contiennent des régions de similarité ou des motifs de séquence similaires dans leur contexte de séquence plus large* » (H. Gomaa and A. Fahmy, 2013).

Cet algorithme a été élaboré au départ pour estimer la similarité entre des séquences d'acides aminés dans l'ARN, mais il a été « exporté » vers d'autres disciplines, le site de l'université de Fribourg propose un tutoriel pour comprendre son fonctionnement et effectuer les calculs[156].

Nous ajoutons que cet algorithme est principalement utilisé pour vérifier les variations de type *perte de mots insignifiants, petits changements* et dans une moindre mesure *ordre des mots,* c'est-à-dire des changements, souvent à l'échelle du caractère dans un « mot » ou dans une petite séquence de caractères[157]. Nous voulons dire qu'il donne de très bons résultats de similarité pour les deux premières variations et de moins bons pour la troisième. Mais comme nous travaillons sur de courtes séquences (quelques mots) et comme nous supposons le principe de compositionnalité respecté, nous pensons que l'ordre des mots n'affectera pas de façon trop importante les séquences étudiées, nous verrons dans la suite comment nous traitons ce positionnement. Comme c'est un algorithme qui opère sur la séquence, il est aussi totalement affranchi du langage sur lequel il travaille[158].

L'algorithme mesure la proximité relative et donne un score maximal pour les deux séquences les plus proches parmi l'ensemble des séquences comparées, en fonction de trois paramètres, un qui augmente le score et deux qui le diminuent :

- *match*, la correspondance exacte, augmente le score ;

153 Il ne s'agit donc ici en aucun cas de vérifier la similarité sémantique de l'unité analysée.
154 Source : (Silva, 2018).
155 Sources (*Bill Buchanan*, no date).
156 Source : (*Teaching - Smith-Waterman*, no date).
157 C'est le cas des accords en genre et en nombre, de la ponctuation, mais aussi de très nombreuses constructions syntaxiques qui articulent les syntagmes entre eux.
158 Il y a aura un travail de nettoyage des données simple à effectuer : supprimer les ponctuations, les espaces et mettre le texte en minuscule.

- *mismatch*, la non-correspondance, diminue le score ;
- *gap*, l'écart entre deux caractères, diminue le score.

On accordera un poids plus ou moins important à chaque paramètre en fonction du besoin et de la tolérance à l'erreur. Très empiriquement, les valeurs suivantes nous donnent des résultats satisfaisants :

$$v_{match} = +3\ ;\ v_{mismatch} = -3\ ;\ v_{gap} = -1$$

Nous accordons le même poids, mais en sens opposé pour la correspondance et la non-correspondance, une manière de peu tolérer deux caractères différents, mais nous acceptons plus volontiers un écart entre deux caractères dans la même séquence.

Les deux séquences avec le score le plus élevé sont donc celles qui ont le plus de similarité, mais cela ne veut en aucun cas dire qu'elles le sont « suffisamment ». C'est seulement le meilleur résultat possible obtenu dans un jeu de n séquences que l'on compare. Dans notre cas, nous voulons une mesure de similarité absolue entre deux séquences. Pour cela, nous proposons de faire le rapport entre le score obtenu entre les deux séquences avec le score maximum que pourrait virtuellement obtenir la séquence la plus courte.

Le score virtuel maximum est défini tel que :

$\max(score_{virtuel}) = \min(n_{char}) * v_{match}$ (équation 10).

Comme cela, on peut effectivement déterminer le degré de similarité entre deux séquences et voir leur proximité absolue. Il faudra fixer un seuil de significativité de la similarité pour considérer deux instances comme identiques.

Par exemple : nous comparons une séquence de référence *onthejune* avec des séquences « extraites » par notre algorithme *NER* et nous posons :

$$seuil_{significativité} = 0{,}75$$

Nous obtenons les résultats suivants :

séquence référence : onthejune					
séquence à comparer	meilleure séquence	score	nb caractère	score virtuel max	rapport score/score virtuel max
onthejune	onthejune	27	9	27	1,000
onthemay	onthe	15	8	24	0,625
onjune	on___june	15	6	18	0,833
onmay	on	6	5	15	0,400

Tableau 18 : degré de similarité entre une séquence de référence et des séquences candidates

Nous voyons que deux séquences candidates sont très proches de la séquence de référence. Ces deux séquences seront fusionnées tandis que les deux autres nécessiteront la création de nouvelles instances. Ainsi :

- pour les séquences candidates à l'instance de classe *TimeInterval*, nous effectuons une première comparaison des caractères numériques qui permet d'emblée de déterminer si deux séquences sont identiques. Si les caractères

numériques sont identiques, alors on élargit la comparaison à la séquence complète ;

- pour les séquences candidates aux autres instances, on évaluera directement la similarité absolue entre les séquences candidates ;
- en fonction du score obtenu[159], notre algorithme *NER* créera ou non de nouvelles instances dans l'ontologie.

Ci-dessous est une illustration de la gestion des instances en fonction de leur similarité en vue de leur écriture dans l'ontologie est présentée ci-dessous :

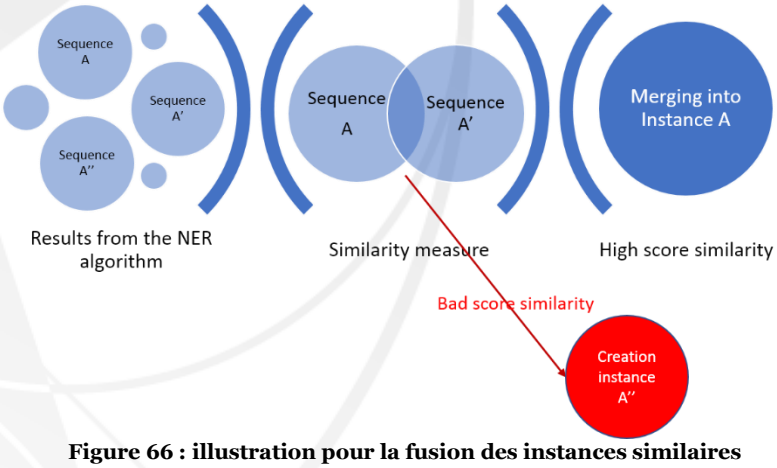


Figure 66 : illustration pour la fusion des instances similaires

Nous avons anticipé deux problèmes, la structure minimale acceptable et la possibilité de doublons, auxquels nous seront probablement confrontés lors du déploiement de notre algorithme *NER*. Nous revenons maintenant à l'exploitation directe de la population de l'ontologie avec la mise en place d'un graphe d'évènements. Pour cela, il faut d'abord dresser une chronologie.

3.1.3.5.2 La chronologie des instances d'évènement

Il existe des algorithmes capables d'effectuer le tri par ordre chronologique à partir de données calendaires[160]. Ce type d'algorithme est capable d'interpréter des séquences alphanumériques comme des expressions de date ou d'horaire (appelés *TimeStamp*) et permet d'effectuer des classements. Nous utiliserons un algorithme de ce type pour ordonner chronologiquement nos instances d'évènement.

Ceci nous permettra également de relier entre elles les instances par les prédicats de temporalité *follows* et *precedes* dans *DOLCE*.

En simulant notre algorithme *NER*, nous obtenons, par ordre d'apparition dans le texte, les phrases d'évènement suivantes :

159 Des essais seront nécessaires pour définir un seuil de similarité à franchir pour considérer les instances comme identiques ou non.

160 Merci à Ragheb Ghandour pour les sources (*A Deeper Look into the Java 8 Date and Time API - DZone Java*, no date; *java - How to sort timestamp list?*, no date; *java - Sort objects in ArrayList by date?*, no date; *python - How do I sort a list of datetime or date objects?*, no date).

1 « *On April 25, as efforts to actuate the BOP stack continued, BP began to consider placing a large containment dome, also known as a cofferdam, over the larger of the two leaks from the broken riser.* »

6 « *Following an MMS inspection of the Discoverer Enterprise, BP began to lower the 98-ton dome to the sea floor late in the evening of May 6.* »

18 « *When crews started to maneuver the cofferdam over the leak at the end of the riser on the evening of May 7, hydrates formed before the dam could be put in place, clogging the opening through which oil was to be funneled.* »

Donc, si l'on extrait les entités temporelles (qui formeront des instances *TimeInterval* dans l'ontologie) et qu'on les classe par ordre chronologique grâce à un algorithme qui gère les *TimeStamp*, on a la séquence temporelle suivante :

$$On\ April\ 25 < of\ May\ 6 < of\ May\ 7$$

Et la chronologie évènementielle est donc la suivante :

$$phrase_1 \rightarrow phrase_6 \rightarrow phrase_{18}$$

Chaque phrase d'évènement a généré des instances de classes *Agent*, *Action* et *Place*. On pourra proposer des « assemblages » d'instances en fonction de ce que l'on souhaite voir apparaître. Par exemple, avec le code suivant :

$$Agent + Action - Place$$

La séquence sera construite avec les instances *Agent* et *Action* et en retirant l'instance *Place* (contenue dans l'instance *Action*). On pourra alors générer le graphe d'évènements suivant :

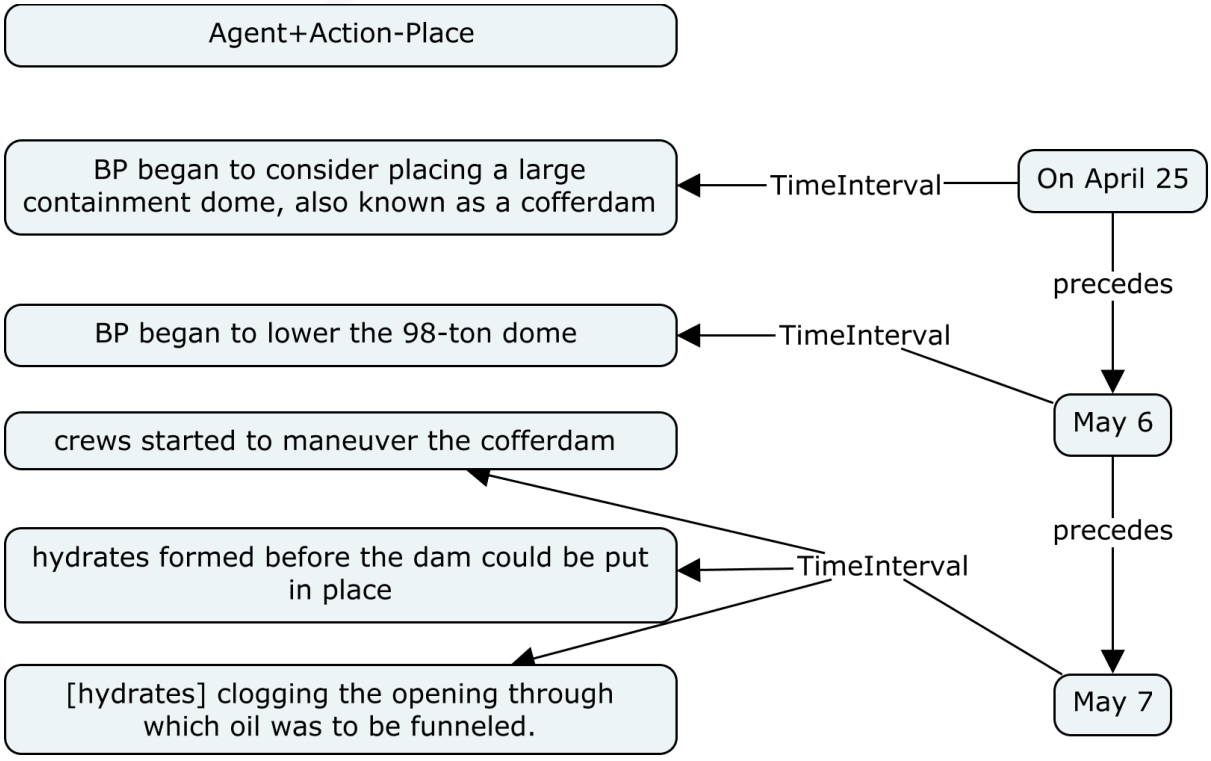


Figure 67 : graphe d'évènements du *cofferdam*

Nous proposons donc au niveau théorique un algorithme *NER* capable de reconnaître des structures syntaxiques d'une phrase en les comparant à un référentiel que nous avons mis au point, le *Csas*. Les résultats obtenus sont ensuite filtrés par un algorithme de similarité binaire afin d'éliminer d'éventuels doublons et les séquences de mots retenus sont écrites dans l'ontologie et en deviennent des instances. Les séquences textuelles exprimant une temporalité sont, quant à elles, détectées par le biais de notre algorithme *NER* de la même manière que les autres séquences, mais elles sont passées, par la suite, au crible d'un algorithme qui est capable d'en générer des *TimeStamp*, c'est-à-dire des marqueurs temporels qui permettent de créer une chronologie entre *TimeStamp* et d'orienter ainsi les instances d'évènement en fonction du temps. Nous sommes donc capables de fournir un graphe de la séquence d'évènements exprimée dans un texte.

Cela nous donne la transition sur la manière d'aborder la causalité par l'événementialité.

3.2 Traitement automatique de la causalité

Dans cette partie, nous allons présenter notre proposition pour la création d'une machine, de type *question answering* capable, à partir d'un questionnement émis en langage naturel par un humain à propos du *pourquoi* de la survenance d'un évènement, et, basée sur l'analyse de données textuelles, d'apporter une réponse par le biais d'une représentation graphique du cheminement causal de l'argumentaire exposé dans le texte.

3.2.1 De la causalité

La causalité est la relation de cause à effet[161]. Comprendre pourquoi un évènement (ou un phénomène) se produit est une quête scientifique majeure dans de très nombreuses disciplines. Comme nous l'avons présenté dans le chapitre 1, la détermination des causes est l'objet premier d'un rapport d'enquêtes à la suite d'un accident et les *safety studies* n'échappent pas à cette « règle ». Nous parlerons de causalité pour décrire l'ensemble des relations de causes à effets (au pluriel) qui permettent de répondre à la question posée par Griffin :

> « *Quelle est l'influence de la cause d'un antécédent temporel sur ce qui s'est produit par la suite dans un évènement ?* » (Notre traduction 1993, p. 1000)

Traiter de la causalité, c'est donc traiter en premier de l'événementialité et donc de la temporalité. Cela justifie l'effort fourni en première partie de ce chapitre pour déterminer des instances d'évènements dans un texte et nous assure la transition vers la détermination de la causalité entre ces évènements.

Notre réflexion à propos de la causalité est bien ancrée sur la théorie de l'*Event Structure Analysis* (*ESA*) amenée par les travaux de David Heise (*Event Structure Analysis*, no date) et Larry Griffin (1993; 1998) et notamment, la mise au point et

161 Définition donnée par le CNRTL (*CAUSALITÉ : Définition de CAUSALITÉ*, no date).

l'utilisation depuis une quinzaine d'années du logiciel *ETHNO* que nous allons voir par la suite (Griffin, 2007).

Au-delà de la théorie qui a permis d'« ouvrir », de « déplier » de nombreux évènements divers et variés[162], nous retenons qu'elle a marié « *la logique et le mode de fonctionnement de l'ESA à des questions d'interprétation causale, de narration, de causalité historique, de méthode comparative, etc.* » (Notre traduction Griffin and Korstad, 1998, p. 164).

L'*ESA* a amené un formalisme structurel, et dans la représentation graphique de l'évènement (son découpage en actions orientées chronologiquement), et dans le questionnement de causalité (dont nous allons longuement parler par la suite), qui apparait indispensable aujourd'hui en termes d'ingénierie des connaissances, particulièrement à l'heure où la possibilité d'exploiter en masse et de façon (presque) automatique des données textuelles pour en tirer des conclusions du même ordre, mais à une échelle « industrielle » et de manière plus robuste. En effet, nous allons le voir par la suite, notre proposition dépasse le logiciel *ETHNO* dans le sens où notre logiciel est capable justement de « répondre à ses propres interrogations » parce qu'il « découvre lui-même » la causalité exprimée dans un texte et permet *in fine* à l'analyste, non pas de réfléchir à ce que pourrait être la causalité d'un évènement, mais plutôt de questionner la causalité telle qu'elle est exprimée. Nous confirmons ici notre volonté d'approche par la donnée de ce que sont les choses et non pas le contraire. C'est au texte de nous dire s'il s'agit de causalité ou non. Pour montrer la portée de notre réflexion, nous allons faire un exemple et analyser avec d'abord avec *ETHNO* un texte qui traite de l'échec de l'installation du *cofferdam*. Nous voulons extraire la structure causale proposée par le texte *B Cofferdam* de 38 phrases, extrait du *swp* n°6 qui relate l'histoire du *cofferdam* et explique son échec. Ce texte sera notre domaine de définition.

3.2.1.1 Une première approche de la causalité dans les narrations : le logiciel ETHNO

Si l'on se réfère aux travaux de Griffin et de Heise, il y a quatre manières de se poser la question à propos d'expressions de la causalité (no date a, p. 16), avec une action (évènement) *A* qui précède chronologiquement une action *B* :

- prérequis : *est-ce que B exige A ou une action similaire ?*
- implication : *l'occurrence de A implique-t-elle B ou une action similaire ?*
- lien de causalité imminent (ou historique) : *est-ce que A ou une action similaire est une cause de B dans les circonstances qui existaient ?*
- contrefactuel : *supposons qu'une action comme A ne se produise pas. Est-ce que B peut se produire de toute façon ?*

Heise précise « *qu'une forme de question peut sembler plus lucide que les autres* » pour aider à cerner le lien de causalité. Toutefois, « *les questions sont logiquement équivalentes. Répondre Oui aux trois premières questions, ou Non à la question*

162 Source : http://www.indiana.edu/~socpsy/ESA/ESApubs.html

contrefactuelle, indique que les deux actions sont liées. La réponse Non aux trois premières questions, ou Oui à la question contrefactuelle, indique que les deux actions ne sont pas liées » (Heise, no date b, p. 16). Toutes amènent le lecteur à se questionner à propos de l'existence d'une relation de causalité entre deux évènements survenus chronologiquement. Il faut préalablement créer une séquence orientée chronologiquement et l'évènement décrit dans l'action 1 doit survenir avant l'évènement dans l'action 2 et ainsi de suite. De plus, une phrase construite avec des structures grammaticales qui amènent une temporalité ou une séquence doit être décomposée en composants élémentaires de cette séquence.

Le logiciel *ETHNO* est un logiciel d'analyse d'évènements dont le matériau support est le texte écrit[163]. L'objectif est de structurer le texte en une succession chronologique d'actions (que nous assimilerons aux évènements) dont on déterminera par la suite s'il y a lien de causalité entre elles grâce à un questionnement proposé à l'analyste. En partant du récit de l'évènement majeur, l'analyste va opérer diverses structurations dans le texte afin d'améliorer sa propre compréhension notamment au sujet de la causalité entre les évènements.

Exemple : supposons un texte qui contient, quel que soit leur position dans ce texte, les phrases suivantes :

A. Il y a eu un défaut de veille sur les deux navires.
B. Le ferry a abordé le porte-conteneurs.

Nous voulons vérifier s'il existe un lien de causalité tel que $A \rightarrow B$.

1. Il faut d'abord considérer un ordre chronologique entre les deux phrases. Soit A puis B.

Puis nous allons choisir la question contrefactuelle pour nous aider à résoudre le lien, cela serait quelque chose comme : s'il n'y avait pas eu de défaut de veille entre les deux navires, le ferry aurait-t-il tout de même abordé le porte-conteneurs ? Soit en logique propositionnelle, nous posons $\neg A \rightarrow B$ et vérifions sa véracité. Admettons que nous répondions OUI, alors nous pouvons affirmer que A n'est pas une cause de B car les deux propositions $A \rightarrow B$ et $\neg A \rightarrow B$ montrent que A n'influence pas B et n'apporte pas d'information. Admettons que nous répondions NON, alors A est une cause nécessaire (indispensable) [164] à la survenance de B.

Nous allons maintenant analyser l'échec du *cofferdam*. Le *cofferdam* est un objet technique dont l'objectif de conception est de collecter et de canaliser l'effluent (un mélange principalement composé de gaz, d'huile et d'eau de mer qui remonte par poussée d'Archimède vers la surface) depuis le lieu de collecte, l'extrémité arrachée du *riser* posée sur le fond de l'eau) jusqu'à un navire de traitement en surface. Pour cela,

163 Source : http://www.indiana.edu/~socpsy/ESA/

164 *ETHNO*, avec ce type de questionnement, amène l'analyste à se poser la question de la nécessité (du caractère indispensable) de la cause. Les quatre questions impliquent que dans tous les cas, fusse l'action A non suffisante, elle n'en est pas moins indispensable (nécessaire) à la survenance de B. C'est le postulat de Heise rappelé par Griffin dans une note de bas de page (1993, p. 1113).

le *cofferdam* est conçu comme un entonnoir inversé (la tête en bas) et équipé d'un flexible à son extrémité supérieure, connecté au navire de traitement. Une fois en position, un navire descendra le *cofferdam* par grutage à l'aplomb de l'extrémité du riser[165] jusqu'au posé sur la fuite en question, au niveau du fond marin. L'objectif est donc de diminuer les hydrocarbures qui s'échappent dans l'environnement en les collectant à la source, laissant un temps de répit à l'organisation d'intervention pour tenter de trouver une solution définitive. Une description synthétique de cet équipement est donnée par Eude et al. (2016, pt. 3.4 The cofferdam). Ci-dessous une photo du *cofferdam* prise au moment de sa mise à l'eau :

Figure 68 : le *cofferdam* mis à l'eau, provenance de la photo : (WashingtonsBlog, n.d.)

Nous choisissons la séquence évènementielle suivante[166] :

1 « *On April 25, as efforts to actuate the BOP stack continued, BP began to consider placing a large containment dome, also known as a cofferdam, over the larger of the two leaks from the broken riser.* »
2 « *By May 4, just ten days after first raising the possibility of using a containment dome, BP reported that it had finished modifying for deep-sea use a preexisting dome that was 14 feet wide, 24 feet long, and 40 feet tall.* »
3 « *Following an MMS inspection of the Discoverer Enterprise, BP began to lower the 98-ton dome to the sea floor late in the evening of May 6.* »

165 Une des deux fuites d'où s'échappent les hydrocarbures du puits.
166. Ces phrases correspondent à l'idée que nous avons de la notion *event sequence* des travaux de Griffin. Notre algorithme *NER* doit pouvoir détecter ce type de phrase.

4 « *When crews started to maneuver the cofferdam over the leak at the end of the riser on the evening of May 7, hydrates formed before the dam could be put in place, clogging the opening through which oil was to be funneled.* »

A partir de ces phrases d'évènements orientées chronologiquement, nous allons questionner l'éventuelle causalité qui lierait les évènements.

Le questionnement se présente sous cette forme :

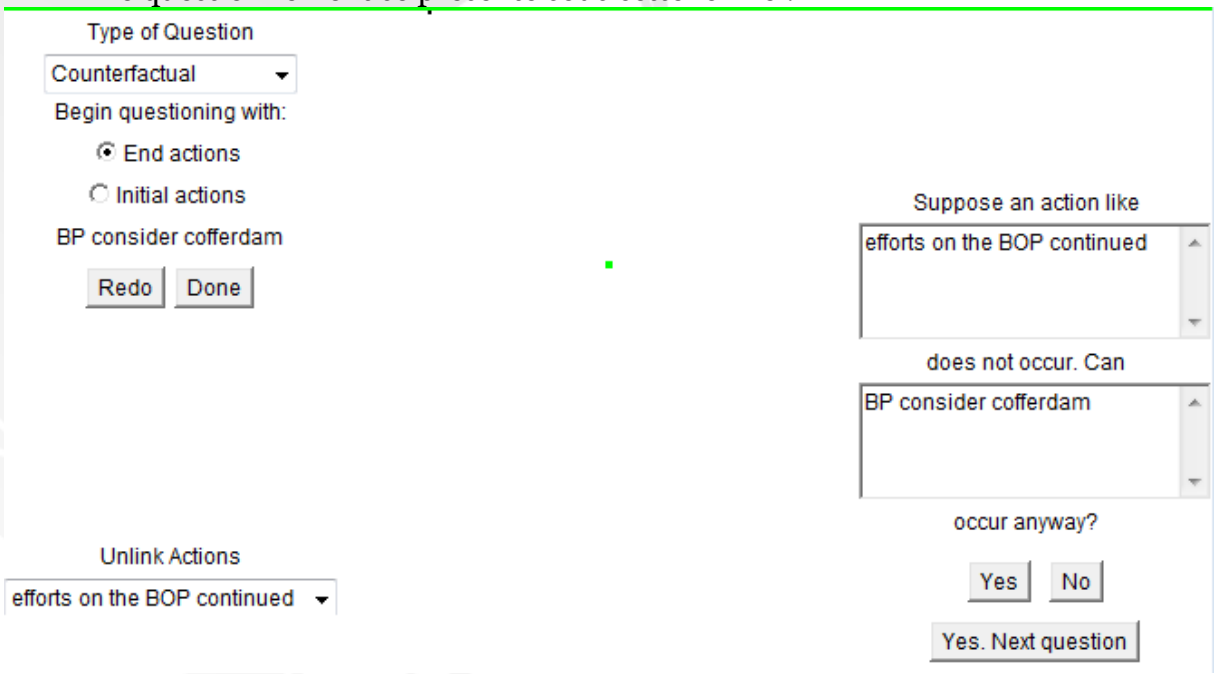


Figure 69 : questionnement causal de type contrefactuel

La réponse à la question posée permet de créer un lien de causalité entre les deux évènements que nous tentons de lier au-delà de leur chronologie. La question contrefactuelle est posée ici de la sorte :

Est-ce-que l'évènement « BP consider cofferdam » surviendrait quand même si l'évènement « efforts on the BOP continued » ne survenait pas ?

Dans cet exemple, si nous répondons *Yes*, alors nous pouvons affirmer que l'évènement « *efforts on the BOP continued* » n'est pas une cause de la survenance de l'évènement « *BP consider cofferdam* ». Au contraire *No* signifie qu'il existe au moins un lien de causalité entre ces deux évènements.

Le logiciel nous permet donc de questionner évènement après évènement la relation de causalité et nous obtenons à la fin un graphe de causalité qui relie les évènements entre eux. Nous obtenons donc le graphe de causalité présenté en page suivante :

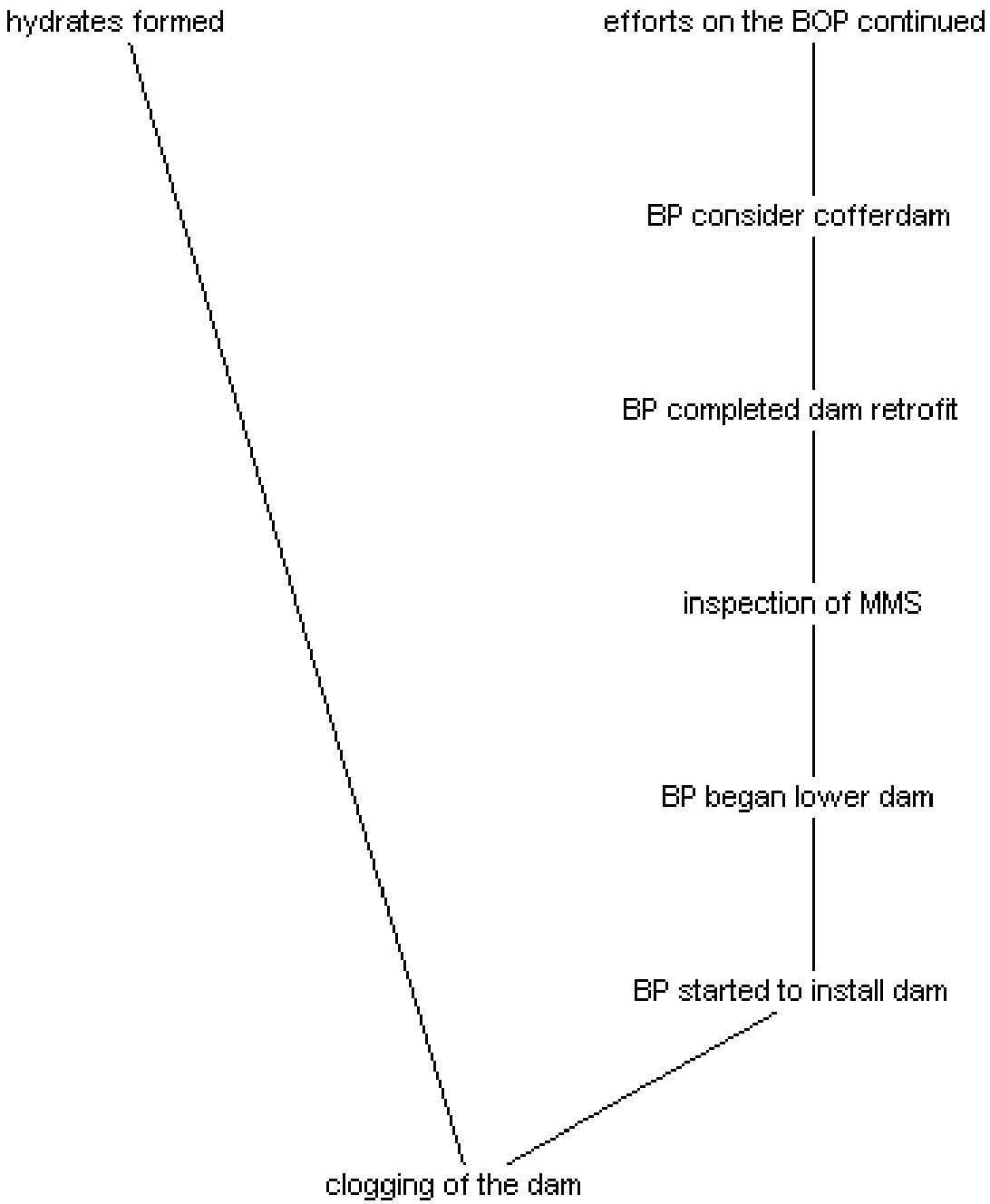


Figure 70 : une analyse de la causalité à l'aide du logiciel *ETHNO*

On comprend avec ce graphique que l'évènement « *clogging of the dam* » est le résultat d'un ensemble d'évènements *précédents* qui ont causé sa survenance. Il y a deux cheminements causaux en parallèle, l'un sur le déroulement de la vie du *cofferdam*, depuis sa « conception » jusqu'à son déploiement ; l'autre qui relie directement l'évènement « *hydrates formed* » avec « *clogging of the dam* ». On précisera que le lien de causalité entre les évènements « *Inspection of MMS* » et « *BP began lower dam* » est établi de manière contrefactuelle tel qu'il n'aurait pas été possible pour BP

de descendre le *cofferdam* sans l'autorisation accordée après inspection par le MMS de la procédure de descente.

On peut conclure que l'évènement « *clogging of the dam* » est la conséquence directe de la survenue de deux évènements, « *hydrates formed* » ET « *efforts on the BOP continued* ».

3.2.1.2 Les limites d'ETHNO

Cependant, lorsqu'on lit le texte d'où l'analyse est tirée, on a en sus des explications à la survenance de l'évènement « *hydrates formed* » qu'on ne peut faire apparaître avec *ETHNO*. En effet, ces explications ne sont pas liées dans le texte à des actions ou des évènements tels que conceptualisés dans les travaux de Heise et Griffin. Typiquement, la formation des hydrates est un phénomène qui se serait produit quels que soient les évènements liés à l'activité du moment (la lutte pour fermer le puits Macondo). Pour le vérifier, on pourrait inviter le lecteur à se renseigner sur la formation des hydrates de gaz en grande profondeur[167], et surtout, on apprend, à la lecture du texte, que les protagonistes (les acteurs de l'action) sont parfaitement conscients de ce phénomène. Pour paraphraser Griffin (1993), la *non-formation* d'hydrates de méthane dans ces circonstances *n'est pas* une alternative plausible contrefactuelle pour expliquer la survenance de l'évènement étudié ; ils se seraient formés « de toute façon ». On ne peut donc pas considérer la formation des hydrates comme l'une des causes de l'échec du *cofferdam*. Il faut pouvoir aller plus loin dans l'analyse de la causalité pour comprendre comment le *cofferdam* a échoué. Nous mettons le doigt ici sur la limite de la méthode *Event Structure Analysis* : une causalité non ancrée dans un évènement ne pourra être analysée par cette méthode. *ETHNO* (et plus largement l'*ESA*) ne traite que des causes qui s'inscrivent dans un schéma de séquence chronologique et une cause « universelle », affranchie de la temporalité, ne pourra être analysée de façon correcte[168]. Avec la lecture historique de l'évènement, on ne peut pas remonter la causalité pour répondre à la question « pourquoi les hydrates se sont-ils formés ? » qui nous permettrait de comprendre pourquoi ces hydrates se sont formés dans ces conditions, dans ce contexte historique, dans cet évènement. *ETHNO* n'aborde pas les *circonstances* qui ont concouru à l'événement, sans qu'il soit possible d'identifier un lien de nécessité entre elles et l'événement.

Si nous revenons à la lecture complète du texte, un humain comprend que c'est la sous-estimation de la possibilité de formation des hydrates, liée à une sous-estimation du débit de fuite, liée elle-même à une croyance à propos du débit en question, qui a entraîné l'échec du *cofferdam*. Ce n'est donc pas la formation d'hydrates la cause de l'échec du *cofferdam*, mais bien la croyance et les hypothèses liées au débit du puits et la succession causale d'évènements qui ont entraîné l'échec de cette solution.

167 Source : (*Methane Hydrate | Department of Energy*, no date).
168 Ou plutôt, qu'une causalité est valable uniquement dans le cadre d'un schéma existentiel pour la description d'un évènement.

On pourrait représenter à l'aide d'un graphique la causalité telle qu'elle peut être perçue dans le texte :

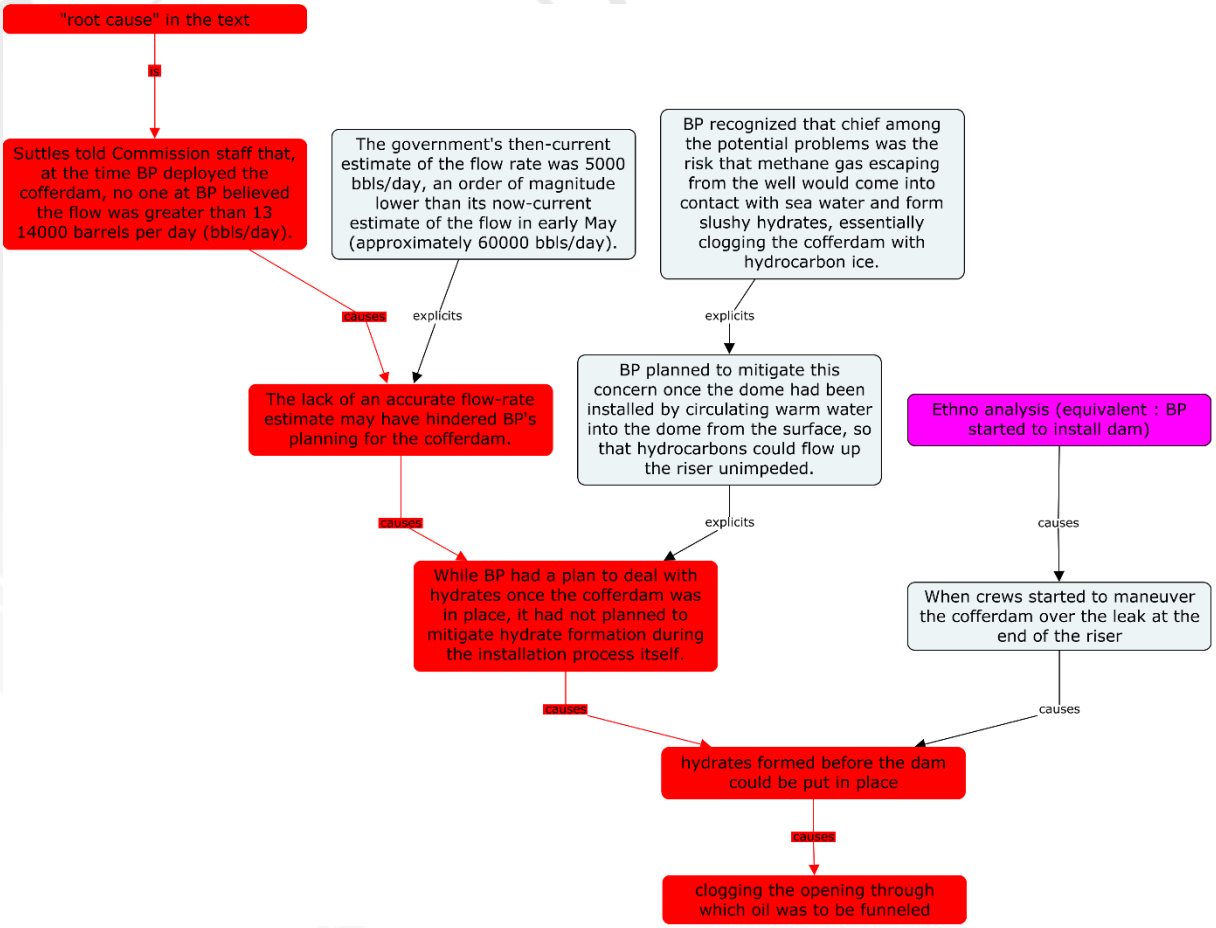


Figure 71 : représentation graphique de la causalité exprimée dans le texte

Ce type de graphe est une façon intéressante d'expliciter la causalité lorsque celle-ci n'est qu'implicite dans le texte. On rejoint l'idée de *template* dans la capacité du graphe à sérialiser les données et proposer une synopse du matériau de recherche (Dumez, 2016). Dans ce cas, le graphe a été dessiné à la main après une lecture attentive du texte et un certain degré d'expertise du domaine considéré. Nous allons voir comment nous pouvons nous approcher de ce résultat par une méthode algorithmique originale.

Nous venons de présenter l'approche « historique » de la causalité comme une succession d'évènements dont la survenance s'explique par des relations causales. Les travaux de Heise et de Griffin ont amené, avec l'*Event Structure Analysis* et le logiciel *ETHNO*, un formalisme, une méthode et un outil pour pouvoir traiter de la causalité dans un texte écrit et en tirer des résultats exploitables. On a vu aussi les limites de cette théorie En 2014, dans leurs travaux d'analyse algorithmique des relations causales et temporelles, Mirza *et* al. (2014, p. 15) ont montré, dans un corpus de texte[169] (peu conséquent, environ 100 000 mots), qu'il y a seulement 33 % de relations causales qui ont un « sous-jacent » temporel (2014, p. 15). De plus, comme nous l'avons montré, *ETHNO* n'est en rien capable de produire des résultats sans une intervention de l'humain à toutes les étapes et ne fournit seulement que les résultats

169 Source : (UzZaman *et al.*, 2013).

programmés par la personne. Griffin dit à ce propos : « *(…) la causalité n'est pas « découverte » par son utilisation. L'analyste, et non le logiciel, possède les connaissances nécessaires pour structurer et interpréter l'évènement* » (2007, p. 6).

Nous allons faire en sorte que la machine soit capable de structurer et d'interpréter l'évènement et la causalité ; la machine devient l'analyste.

3.2.2 Traitement automatique du langage naturel et causalité

Dans le domaine de recherche du TALN, des travaux récents ont été menés pour tenter de faire émerger des schémas linguistiques d'expressions, appelés *patterns,* de la causalité, notamment ceux de Higashinaka and Isozaki (2008) et Rink et al. (2010). Les premiers travaux à ce sujet qui semblent faire référence sont ceux de Girju et al (2003 ; 2002). Les auteurs sont les premiers à dépasser le travail « artisanal » effectué jusque-là dans la recherche d'expressions causales dans un texte en proposant un algorithme capable de traiter à grande échelle un grand corpus de textes. Ils recherchent des *patterns* lexico-syntactiques se référant au lien de causalité et font émerger un ensemble de verbes utilisés à cet effet grâce à un *pattern* intraphrase de type *NP (Cause Noun)-verb-NP (Effect Noun)*. Les résultats font émerger un ensemble de verbes causaux que l'on retrouvera identifié comme tels dans WordNet[170].

Asghar (2016) propose un recensement des moyens algorithmiques d'extraction des relations causales dans un texte écrit en langage naturel. Les travaux qui ont tenté de cerner la causalité exprimée peuvent être classés en deux catégories ; les travaux construits sur une analyse linguistique du texte et ceux qui reposent sur une approche statistique et d'apprentissage. Il existe aussi quelques méthodes d'annotations de texte « à la recherche de la causalité », et qui ont conduit à la création de corpus, comme proposées par Djemaa et al. (2016), Dunietz et al. (2015, 2017b, 2017a), Higashinaka and Isozaki (2008) et Mirza et al. (2014). Une autre approche de la causalité est proposée par Luo et al. (2016) avec la création « *d'un réseau de relations de causalité pondérées avec des co-occurrences de causalité entre les termes qui est extrait d'un grand corpus web.* » (Luo *et al.*, 2016, p. 422).

Nous reviendrons ici sur ces approches. Mais en préliminaire, testons l'un des algorithmes jugé le plus performant aujourd'hui dans le domaine.

3.2.2.1 Allen NLP

Nous avons choisi d'utiliser *Allen NLP*[171] qui est une Intelligence Artificielle directement accessible en ligne sans besoin de connaissance informatique pour l'implémenter ou l'utiliser. Le code est open-source et disponible sur la plateforme GitHub[172]. Le porteur du projet est l'*Allen Institute for Artificial Intelligence*[173], fondé par Paul Allen, le co-fondateur de Microsoft.

Nous allons maintenant nous intéresser en particulier à la machine *question answering* de cette IA. Cette machine est bâtie sur un algorithme proposé par Gardner

170 Nous présenterons la sémantique et grammaire de la cause par la suite.
171 Source : http://allennlp.org/
172 Source : https://github.com/allenai/allennlp
173 Source : http://allenai.org/

et al. (2018) sur la base des travaux de Peters et al. (2018). La machine a obtenu de très bons résultats face à un benchmark, *The Stanford Question Answering Dataset*, (*SQuAD)*. Ce *benchmark* est lui aussi très récent et semble se positionner comme une référence dans cette branche de l'IA au service du TALN (Rajpurkar et al., 2016 ; « The Stanford Question Answering Dataset, » n.d.)[174].

Nous soumettons à l'IA le même texte que nous avons étudié avec *ETHNO*, la partie *B Cofferdam*, extrait du document *staff working paper n°6* qui relate l'histoire du *cofferdam* et qui comporte 38 phrases. Puis, nous posons à l'IA la question suivante (la question-test) :

Figure 72 : question à l'IA

Nous obtenons la réponse suivante :

Figure 73 : réponse de l'IA

Cela peut paraître tout à fait satisfaisant. En effet, c'est bien la cause que nous avons déterminée avec *ETHNO* même si on a vu que ce n'était pas suffisant.

Mais si l'on regarde de plus près, on accède au contexte considéré par la machine pour comprendre « sa décision » :

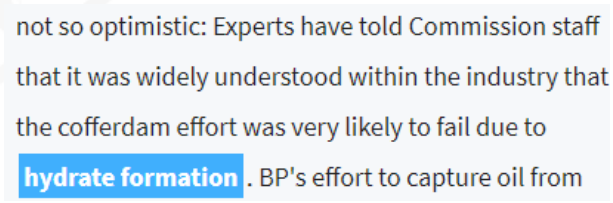


Figure 74 : réponse contextualisée dans le texte

Cette expression est surlignée uniquement dans cette phrase (n°15). Or, à la lecture de cet extrait, on comprend que l'expression de cet échec est hypothétique. C'est-à-dire que l'expression de cette causalité est une projection d'idée, une alternative plausible proposée *a priori* (Griffin, 1993). Mais elle n'est pas l'expression de cet échec. En fait, la machine ne commet pas d'erreur dans sa réponse, mais sa justification n'est pas la meilleure. De plus, comme nous l'avons vu à la lecture, le cheminement causal dans le texte est tout de même plus profond. Mais nuançons notre évaluation, car nous l'avons vu aussi, il n'y a pas de « meilleure phrase » qui puisse exprimer la raison de l'échec du *cofferdam* que celle identifiée. En effet, il n'y a aucune phrase qui ressemblerait peu ou

174 « Les modèles de « Machine Comprehension » (MC) répondent aux questions en langage naturel en sélectionnant une plage de réponse dans un texte de données probantes [preuves]. Le modèle *Allen NLP* MC est une ré implémentation de l'algorithme *BiDAF* ou « *Bi-Directional Attention Flow* » (*Bidirectional Attention Flow for Machine Comprehension - Semantic Scholar*, no date; Seo *et al.*, 2016), une référence dans le MC largement utilisée et qui a atteint une précision aigüe sur l'ensemble de données SQuAD en 2017 » (*AllenNLP - Models*, no date) Le site http://demo.allennlp.org/machine-comprehension/ héberge l'interface utilisateur.

prou à « *the cofferdam has failed because of the hydrate formation.* » Nous allons effectuer quelques tests sur le même texte et commenter les résultats.

Nous posons des questions dont l'objet tourne autour de la réussite ou de l'échec du *cofferdam ;* de la détermination des causes, de l'évènement ou encore du phénomène de formation des hydrates, la réponse à la question précédente. La justesse de la réponse est évaluée par nos soins compte-tenu de notre connaissance du cas.

N°	Human Question	AI Answer	Justesse
1	Why has the cofferdam failed ?	hydrate formation	VRAI
2	Why was the cofferdam a failure ?	hydrate formation	Partiellement VRAI
3	Why was the cofferdam a success ?	the cofferdam did not risk damaging the well or otherwise making the spill worse	Partiellement VRAI
4	Was the cofferdam a failure ?	hydrate formation	Partiellement VRAI
5	When did the cofferdam fail ?	April 25	FAUX
6	Why did hydrates form ?	they also rendered the containment dome buoyant as it was still being lowered	FAUX
7	How did hydrates form ?	lighter than water	FAUX
8	Why do hydrates form ?	they also rendered the containment dome buoyant as it was still being lowered	FAUX

Tableau 19 : jeu de questions et évaluation de la justesse des réponses

Les résultats sont globalement peu satisfaisants.

On remarque que la question 1, très similaire à la question-test (seule la conjugaison change, la temporalité, le passé, reste identique,) donne un résultat juste.

Le résultat 2 est partiellement vrai car il cible en effet une cause qui a amené à l'échec, mais il ne répond pas à la question qui attend plutôt comme réponse un effet de la formation d'hydrates sur le *cofferdam* pour le considérer comme un échec. Typiquement l'extrait suivant de la phrase 18 du texte aurait été plus pertinent :

> « *[…] hydrates formed before the dam could be put in place, clogging the opening through which oil was to be funneled.* »

La réponse aurait pu être : « *clogging the opening* ». La question 3 est une sorte de « question-piège » dont on n'attend pas forcément de réponse , mais le résultat proposé est intéressant car même s'il ne remplit pas entièrement la qualité de ce qui devrait être un succès (et pour cause, il n'y a pas d'éléments dans le texte qui considère le *cofferdam* comme un succès), on voit que la réponse proposée a du sens : il y a une connotation positive associée à la sémantique de la réponse, qui semble donc être considéré dans le même espace sémantique que celui du succès.

La question 4 attend une réponse binaire (que nous savons impossible de donner par l'algorithme compte-tenu du texte soumis) et la machine nous ramène à ce qui fait une des causes de l'échec. Il n'y a donc pas d'interprétation « globale » de la lecture du texte.

La question 5 est d'une nature différente parce qu'elle attend une expression de temps comme réponse. C'est le cas, mais la réponse est fausse ; et pourtant, la réponse pourrait être dans le texte[175], à la phrase 18 :

> *« When crews started to maneuver the cofferdam over the leak at the end of the riser on the evening of May 7, hydrates formed before the dam could be put in place, clogging the opening through which oil was to be funneled. »*

L'ensemble de la lecture de la phrase fait comprendre à un humain que l'obstruction du *cofferdam* a eu lieu le 7 mai.

Ici, nous mettons le doigt sur un point fondamental pour la construction de la machine de *question answering* basée sur la sémantique au sens large : l'intension et les extensions sémantiques des lemmes du lexique. Nous supposons que cette machine « ne sait pas » ce que signifie la forme *clogging* et peut-être le lemme *clog*. Une autre hypothèse est qu'il n'y a pas d'induction de connotation négative dans l'espace sémantique de *clog* par l'espace co-textuel[176] ou enfin, si c'est pourtant le cas, que la machine ne le prenne pas en compte. Nous pensons le confirmer avec ce qui vient.

Les questions 6, 7 et 8 tentent de « remonter le fil » de la causalité pour obtenir des réponses concernant le phénomène de formation des hydrates. Les trois réponses proposées ne conviennent pas. Dans ce cas présent, il y a au moins une réponse explicite, surlignée en rouge, à la formation des hydrates qui est contenue dans la phrase 12 :

> « *BP recognized that chief among the potential problems was the risk that methane gas escaping from the well would come into contact with sea water and form slushy hydrates, essentially clogging the cofferdam with hydrocarbon ice.* »

La machine ne l'a pas vu. De plus, la phrase 12 contient la même forme que la phrase 18, *clogging*, pour l'expression du lemme *clog*. La phrase 12 l'exprime comme un risque, la phrase 18 comme la réalisation « effective » de ce risque, elle l'instancie comme un évènement. Il y a donc deux expressions du type :

$$hydrate_{formation} \rightarrow clogging_{cofferdam}$$

Mais la machine ne le saisit pas. Cela rejoint notre analyse précédente qui montre que *clogging*, même sans connotation négative, possède une signification (une dénotation) qui ne semble pas, pour la machine, être nécessaire pour l'arbitrage entre l'échec et la réussite lorsqu'on lui pose la question. Concrètement, elle n'a pas « compris » que l'obstruction du *cofferdam* (par quoi que ce soit) serait la cause de l'échec de ce dernier.

175 Voir le cheminement causal amené précédemment.
176 Une rapide analyse de la co-textualité de « *clog* » (*POS verb*), à l'aide de Sketch Engine dans le corpus *enTenTen15* n'a pas franchement fait ressortir de connotation négative ou liée à l'échec. Voir le lien : http://ske.li/clog_word_sketch.

Nous sommes loin des résultats obtenus par le modèle sur le benchmark *SQuAD* :

Rank	Model	EM	F1
8 Nov 17, 2017	BiDAF + Self Attention + ELMo (ensemble) *Allen Institute for Artificial Intelligence*	81.003	87.432

Figure 75 : score du modèle « ensemble » de l'IA *Allen NLP* sur le benchmark *SQuAD*

Un score *Exact Match* (*EM*) à 81 % (la réponse coïncide parfaitement avec le *span* de texte attendu) et un score F_1 à 87,4 %.

Au vu des résultats peu satisfaisants, nous avons l'intuition de changer notre questionnement en utilisant l'adverbe *How* en place de *Why*. *How* est utilisé pour introduire un questionnement lié à la manière de..., le plus souvent traduit par Comment... ? en français.

Nous posons deux questions à propos du phénomène d'obstruction du *cofferdam* ; voici les réponses :

Human Question	AI Answer	Justesse
Why did the cofferdam get clogged?	the cofferdam did not risk damaging the well or otherwise making the spill worse	FAUX
How did the cofferdam get clogged?	with hydrocarbon ice	VRAI

Tableau 20 : réponses aux questions *How ?* et *Why ?*

La réponse à la question avec *Why*... est fausse. La réponse à la question avec *How*... est juste, mais elle n'est pas vraie en tant que telle (par rapport à la forme de la question posée). Nous supposons que la préposition *with* a aidé à formuler la réponse. Encore qu'il faille aussi noter que *with* peut être utilisé en place de *because of* dans certaines phrases[177] lui donnant ainsi une vertu d'explication causale. La « glace carbonique » est pourtant une cause (existentielle) de l'obstruction, mais l'expression *with hydrocarbon ice* peut tout à fait être comprise comme une manière d'obstruer.

On peut le montrer avec l'analyse grammaticale de la phrase :

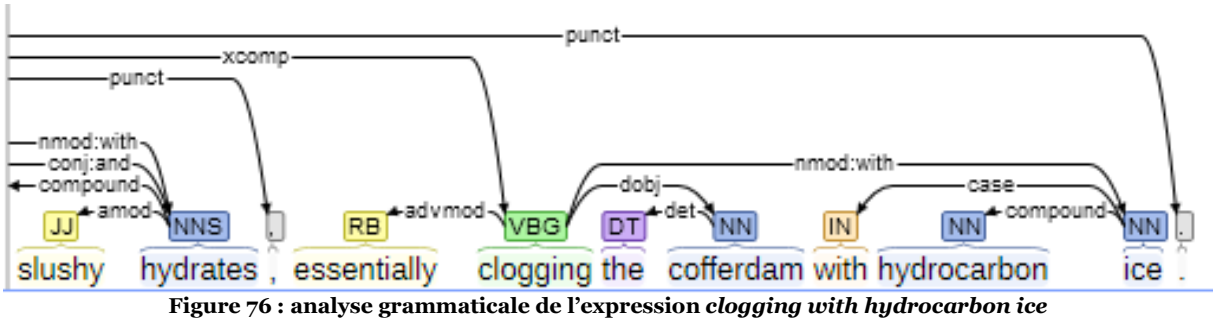


Figure 76 : analyse grammaticale de l'expression *clogging with hydrocarbon ice*

La relation entre *clogging* et *with hydrocarbon ice* est une relation *nmod : with,* elle modifie le nom duquel elle dépend[178].

177 Source : (*With - English Grammar Today - Cambridge Dictionary*, no date).
178 La relation *nmod* est utilisée pour les nominaux dépendants d'un autre nom ou d'un groupe nominal et correspond aux fonctions d'attribut ou de génitif.

Nous mettons ici le doigt sur une différence importante, mais fondamentale entre l'existence de la cause et l'expression de celle-ci dans un texte. En sus de la « réalité » de la causalité effectivement déterminée dans le cas d'un quelconque évènement et déjà difficilement décelable comme nous venons de le voir avec ces quelques exemples, on voit en quoi son expression dans le texte écrit peut amener aussi à un questionnement qui dépasse le cadre théorique de cette thèse, mais dont il faut avoir bien conscience pour la suite des travaux.

En conclusion, nous voyons que la réponse à une question, et notamment à la question *pourquoi*, est tout sauf une chose aisée même pour une intelligence artificielle de haut niveau. Les approximations sont encore trop nombreuses pour estimer le résultat suffisant. Cependant, c'est la piste à moyen terme à privilégier en matière de proposition technologique. En ce sens, développer notre logiciel comme une forme spécialisée de ce type d'IA pour répondre aux questions de temporalité et de causalité nous semble être tout à fait à propos. On pourrait voir notre logiciel comme une forme de module qui viendrait se greffer à une machine *question answering* ; ou voir plus encore en amont, comment la méthode que nous allons présenter puisse par exemple travailler sur le benchmark de l'université de Stanford pour renforcer un modèle d'IA existant (celui d'Allen ou un autre). Il s'agit de développements qui nous attendent à la suite de ces travaux, mais que nous n'aborderons pas ici tant le travail sur la méthode est déjà considérable.

3.2.2.2 Sémantique et grammaire de la cause

Tous les algorithmes d'extraction de causalité reposent sur des marqueurs. Pour identifier les marqueurs sémantiques, nous utilisons *WordNet* que nous avons déjà présenté et nous obtenons les résultats suivants pour la recherche de *cause*. Nous présentons ici les premiers résultats pour chaque catégorie.

- pour le nom *cause* :
    - cause -- (events that provide the generative force that is the origin of something);
    - cause, reason, grounds -- (a justification for something existing or happening));
    - causal agent, cause, causal agency -- (any entity that produces an effect or is responsible for events or results).
- pour le verbe *cause* :
    - cause, do, make -- (give rise to ; cause to happen or occur, not always intentionally);
    - induce, stimulate, cause, have, get, make -- (cause to do ; cause to act in a specified manner).

Nous allons maintenant élargir l'arborescence lexicale autour du nom et du verbe :

- des hyponymes[179] du nom *cause* :

179 Voir définition donnée par le CNRTL (*HYPONYMIE : Définition de HYPONYMIE*, no date).

    - antecedent (a preceding occurrence or cause or event);
    - factor (anything that contributes causally to a result);
    - producer (something that produces).
- des troponymes[180] du verbe *cause* :
    - determine, shape, mold, influence, regulate (shape or influence; give direction to);
    - initiate, pioneer (take the lead or initiative in ; participate in the development of);
    - effect, effectuate, set up (produce);
    - occasion (give occasion to) ;
    - provoke, evoke, call forth, kick up (evoke or provoke to appear or occur);
    - engender, breed, spawn (call forth);
    - motivate, actuate, propel, move, prompt, incite (give an incentive for action);
    - impel, force (urge or force (a person) to an action; constrain or motivate);
    - facilitate (increase the likelihood of (a response)).

L'ensemble de ces unités va former notre champ sémantique de recherche pour l'identification de phrases qui expriment la causalité. De la même manière, nous avons cherché les éléments sémantiques de la conséquence et nous trouvons :

- *consequence, effect, outcome, result, event, issue, upshot -- (a phenomenon that follows and is caused by some previous phenomenon;*
- *consequence, aftermath -- the outcome of an event especially as relative to an individual;*
- *consequence, import, moment -- having important effects or influence.*

Le verbe *effect* renvoie très majoritairement au verbe *cause* que nous avons déjà étudié.

Il existe d'autres référentiels structurels sémantiques, notamment *FrameNet* et *VerbNet*[181] qui ont été créés et sont maintenant unis sous la bannière de l'université Boulder du Colorado dans l'index unifié des verbes où les structures sémantiques sont construites à partir du verbe (« Unified Verb Index, » n.d.). Ce sont des « concurrents » de *WordNet*.

Nous avons construit très rapidement un thésaurus du domaine de la causalité qui nous permet d'avoir l'œil aiguisé pour la recherche d'expressions de la causalité dans le texte. Où l'on voit que la sémantique associée à la causalité est riche et qu'il existe de nombreux verbes et noms pour l'exprimer.

Voyons maintenant les constructions grammaticales de la phrase causale en anglais. Nous cherchons les structures grammaticales qui permettent d'exprimer la

180 Source :: *troponym A verb expressing a specific manner elaboration of another verb. X is a troponym of Y if to X is to Y in some manner* (*wngloss(7WN) | WordNet*, no date).
181 Sources (*LU index | fndrupal*, no date; *Welcome to FrameNet! | fndrupal*, no date; *Verb Sense Annotation Project*, no date; *The Proposition Bank (PropBank)*, no date).

causalité dans l'écriture de la langue anglaise. La structure la plus commune est une phrase constituée de deux propositions, une principale et une proposition subordonnée relative introduite par la conjonction de subordination *because*. Cette conjonction amène la raison d'existence de la proposition principale. Elle en donne l'explication. La proposition principale est donc le résultat (ce qui résulte) de la cause amenée par la proposition subordonnée. Il semble aussi que l'usage de *because* soit quasi exclusif à l'expression de la causalité, ce qui en fait un « marqueur » très sûr de l'expression de la causalité dans un texte même s'il peut y avoir à discussion sur des cas très précis (4.1 Basic Features of Annotations Dunietz, Levin and Carbonell, 2017b)[182]. Il y a également deux autres conjonctions qui sont souvent utilisées : *since* et *as*. Elles ne sont pas exclusives, mais peuvent introduire une causalité. De la même manière la structure grammaticale, *if → then* peut amener une causalité hypothétique. Enfin, la conjonction *so* introduit des clauses de résultat ou de décision qui peuvent être liées parfois à une expression de la causalité.

Même s'il existe quelques marqueurs syntaxiques assez fiables, nous constatons qu'il n'y a pas de catégorie ou de structure grammaticale nécessaire et suffisante dédiée à l'expression de la causalité.

La causalité exprimée est un subtil mélange de sémantique et de syntaxe pour la construction des expressions écrites et d'interprétation pour leur compréhension. Nous allons voir maintenant comment il est possible de cerner cette causalité dans le texte de façon (quasi) automatique.

3.2.2.3 L'annotation de la causalité

Annoter un texte consiste à l'enrichir par le biais d'ajouts de références d'interprétation à partir d'un cadre théorique pour améliorer la compréhension. Annoter correctement un texte est une tâche fastidieuse et méticuleuse qui nécessite aujourd'hui des logiciels adaptés[183]. Nous allons nous inspirer des travaux de Dunietz et al. (2015, 2017b). L'objectif est « *d'annoter toute forme de langage causal – des mécanismes linguistiques conventionnels utilisés qui amènent la cause et l'effet. Ainsi, le schéma ne s'intéresse pas aux relations de cause à effet « du monde réel », mais plutôt aux relations exprimées dans le texte.* » *(Partie : 2.1 BECauSE 1.0 Dunietz et al., 2017a).*

En reprenant l'exemple précédent à propos de l'échec du *cofferdam*, nous allons annoter quelques phrases du texte et en tirer des résultats d'interprétation.

La première étape est l'identification des marqueurs causaux, les *causal connectives* (Partie : 4.1 Basic Features of Annotations Dunietz et al., 2017a) dans la

182 Nous comprenons ces cas comme étant des propositions causales évidentes telles que : « je me suis brûlé avec le feu », une forme de truisme, de lapalissade.

183 Pour effectuer cette annotation sémantique, nous avons besoin d'un logiciel spécifique et le moins que l'on puisse dire, c'est qu'il n'en existe que très peu, la discipline étant particulièrement récente, et nous avons mis un temps certain pour pouvoir en trouver un. Nous sommes partis encore une fois du site du laboratoire *NLP* de l'Université de Stanford (qui utilise le logiciel Brat) et avons alors fini par découvrir (très heureusement) Webanno (Eckart de Castilho *et al.*, 2014), construit à partir de Brat, parfaitement compatible avec notre plateforme et très riche en fonctionnalités.

phrase à annoter. On comprend que la causalité est exclusivement explicite (comme présentée dans le nom du corpus) ; les *causal connectives* servant de marqueurs syntaxiques pour la détection des relations causales (Dunietz, Levin and Carbonell, 2017b, para. 4.1 Basic Features of Annotations). Les auteurs précisent cependant que ces « *proxys* » n'ont pas vocation à être interprétables par une machine bien qu'ils finissent malgré tout par utiliser les *causal connectives* dans leur proposition d'étiquetage automatique de la causalité (3 The Causal Language Tagging Task Dunietz, Levin and Carbonell, 2017a, p. 120).

Si l'on s'en tient à la méthode, voilà comment nous pouvons annoter la phrase suivante :

> *7* « *BP planned to stage a second cofferdam on the sea floor in case the first dam failed.* »

On a :

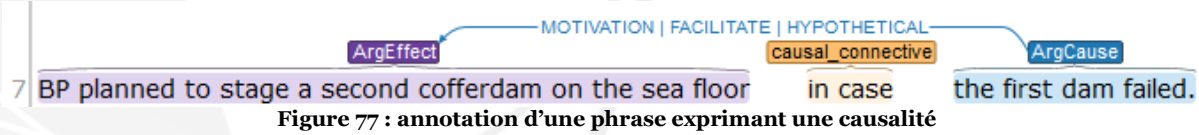


Figure 77 : annotation d'une phrase exprimant une causalité

Le *causal connective* est : *in case*. C'est une forme d'implication ; y existe si et seulement x existe.

Par ailleurs, on a comme arguments de cause et d'effet :

- *ArgCause* est : [the first dam failed.];
- *ArgEffect* est : [BP planned to stage a second cofferdam on the sea floor].

La relation de causalité entre chaque argument est toujours orientée de la cause vers l'effet :

$$ArgCause \xrightarrow{MOTIVATION\_FACILITATE\_HYPOTHETICAL} ArgEffect$$

- *MOTIVATION* car *BP* (qui est l'agent) « perçoit la cause, et donc consciemment pense, ressent ou choisit quelque chose. »
- *FACILITATE* car l'argument cause a la même « polarité » que l'effet produit, le fonctionnement est analogue entre les deux arguments.
- *HYPOTHETICAL* car l'argument cause projette un scénario alternatif, en l'occurrence imaginé par *BP*.

Les auteurs proposent des tests linguistiques pour pouvoir s'assurer de la causalité de la relation entre les deux arguments. Ces tests proviennent des travaux de Grivaz (2010). L'auteur introduit et détaille différents types de tests susceptibles d'aider à établir et obtenir le consensus entre humains à propos de la causalité exprimée dans les phrases. Ces tests sont tout à fait fondamentaux pour nous aider à déterminer la causalité d'une phrase. Même s'il semble qu'ils soient très intuitifs, la formalisation apportée par ces tests nous aidera à questionner cette causalité. De plus, chaque test s'applique sur l'articulation $causal_{connective}(ArgCause, ArgEffect)$, le domaine de définition local, et ce, nonobstant la connaissance que l'on pourrait avoir à propos d'éléments en présence dans le reste du document, mais qui ne seraient pas

exprimés dans la phrase en question. En reprenant la formulation condensée de ces tests des travaux de Dunietz et al. (2017b, para. 4.2 Annotating Overlapping Relations), nous appliquons ces tests à la phrase que nous étudions à l'instant :

Le test du « pourquoi » : après avoir lu la phrase, pourrait-on raisonnablement s'attendre à ce qu'un lecteur réponde à une question « pourquoi » au sujet du potentiel argument effet [*ArgEffect*] ? Si ce n'est pas le cas, ce n'est pas causal.

La réponse est oui : l'argument effet correspondrait bien à une réponse à la question pourquoi BP prévoyait de mettre en place un deuxième *cofferdam* ;

Le test de l'ordre temporel : est-ce que la cause invoquée précède l'effet ? Si ce n'est pas le cas, ce n'est pas causal.

Oui, la cause (hypothétique) précède bien l'effet (potentiel) ; il y a un ordre explicite à la séquence et le verbe *plan* a une dénotation temporelle de projection. L'éventuel montage du deuxième *cofferdam* arrivera de toute façon après l'échec du premier (*a fortiori* du premier) ;

Le test de contrefactualité : l'effet aurait-il été tout aussi probable ou non si la cause ne s'était pas produite ? Si tel est le cas, il n'est pas causal. On a ici :

$$in\ case\ the\ first\ dam\ failed \rightarrow BP\ planned\ to\ stage\ a\ second\ cofferdam$$

On pose :

$$A \rightarrow B$$

avec

$$A = the\ first\ dam\ may\ fail$$

et

$$B = BP\ planned\ to\ stage\ a\ second\ cofferdam$$

Ce qui donne en français une phrase de la sorte : « Le fait que le premier *cofferdam* pourrait échouer a entraîné que BP a prévu le déploiement d'un deuxième *cofferdam* ». Nous allons donc tester l'hypothèse contrefactuelle (le monde plausible selon Griffin) que nous exprimons de la sorte :

$$\neg A \rightarrow B$$

Si l'hypothèse est validée, on en déduit que A n'est pas une cause de B. Si l'hypothèse est invalidée, alors A est bien une cause de B. Ce qui revient à se demander : « Est-ce que BP prévoirait un deuxième *cofferdam* si le premier *cofferdam* ne pouvait jamais faillir ? » La réponse est (très probablement) non[184]. Donc l'hypothèse est invalidée donc nous pouvons en déduire que A est bien une cause de B soit que l'éventualité de défaillance du *cofferdam* est une cause de la préparation d'un deuxième *cofferdam* par BP.

184 Il n'y a rien dans le domaine de définition qui nous permette d'affirmer le contraire (et même largement au-delà de ce domaine car nous n'avons pas vu d'information contraire dans les autres documents notamment présentés dans le chapitre 1).

Le test d'asymétrie ontologique : pourriez-vous tout aussi facilement prétendre que la cause et l'effet sont inversés ? Si tel est le cas, il n'est pas causal. Dans ce cas, il faudrait écrire quelque chose comme : la prévision du second *cofferdam* serait la cause de l'hypothétique faillite du premier *cofferdam*. Cela n'a ontologiquement aucun sens. On ne peut donc pas inverser les arguments.

Le test linguistique : la phrase, peut-elle être reformulée comme suit : « C'est à cause de *X* que *Y* » ou « *X* cause *Y* ? » Si c'est le cas, il est probable qu'il y ait un lien de causalité. Très intuitivement, on voit bien que l'on peut écrire sans difficulté quelque chose comme : l'éventuelle faillite du premier *cofferdam* cause la prévision d'un second par BP. La phrase annotée est donc bien une phrase ontologiquement causale.

La méthode d'annotation proposée par Dunietz et *al.* permet une certaine formalisation par la structuration de l'expression causale autour du *causal connective* et une qualification de ces expressions de causalité *via* la nomenclature qu'ils proposent. Les tests de causalité montrent leur grande importance pour déterminer avec précision la nature de cette causalité et valider les premières intuitions de lecture.

Cependant, différentes limites à leur méthode d'annotation, tant « manuelle » que algorithmique, sont pointées (2017a, p. 119) :

- relations de cause à effet sans déclencheur lexical (« lexical trigger »), [sans *causal connective* identifié] ;
- expressions causales (*connectives*) qui incorporent un moyen ou un résultat [les expressions métonymiques] ;
- les liens qui affirment une relation de cause à effet non spécifiée ;
- les relations comme l'ordre temporel sont souvent réaffectées pour exprimer la causalité ;
- de plus, pour des raisons pratiques, les arguments ne sont annotés que lorsqu'ils apparaissent dans la même phrase que la conjonction.

Cette méthode ne permet donc pas de relier des expressions causales entre des phrases différentes. Les *causal connectives* ne sont seulement de bons repères qu'« à l'intérieur » d'une phrase ; d'après notre compréhension, il s'agit principalement de morphèmes lexicaux qui articulent les propositions entre elles dans une phrase. La causalité qui pourrait être induite par le « passage » d'une phrase à l'autre ne peut donc émerger aux yeux de l'annotateur. Ensuite, la causalité induite dans une phrase telle que « *Il pleut, le sol est mouillé.* » ne saurait être considérée par la méthode d'annotation. Chaque syntagme a une signification qui lui est propre et en fait, pris à part, chaque syntagme est une proposition complète et autonome. Avec une phrase de la sorte, pour le moment, seul un annotateur humain est capable de discerner la causalité induite dans cette phrase. C'est l'association des deux propositions qui génère la causalité en l'induisant. Mais cette causalité aurait pu être tout à fait vraie écrite de la sorte : « *Il pleut. Le sol est mouillé.* », encore une fois, sémantique et consécution des deux phrases induisent la relation de causalité.

Si l'on prend maintenant la phrase suivante :

18 « *When crews started to maneuver the cofferdam over the leak at the end of the riser on the evening of May 7, hydrates formed before the dam could be put in place, clogging the opening through which oil was to be funneled.* »

Cette phrase est un cas intéressant car elle ne possède ni *causal connective* ni même une structure grammaticale explicite de la causalité et pourtant il y a bien une expression de la causalité qui se dessinerait de la sorte :

$$clogging\ the\ opening \leftarrow hydrates\ formed \leftarrow manoeuver\ the\ cofferdam$$

L'annotation ne peut révéler cette causalité. Nous le savions déjà, mais ce qui est important est le caractère « majoritairement » sémantique de cette causalité, qui dépasse la seule expression de la séquence temporelle portée par la structure grammaticale (« *When…, …* »). Les formes *cofferdam*, *hydrates* et *clogging* sont liés par leurs extensions sémantiques, apportées par le contexte (le domaine de définition). Changer n'importe laquelle de ces formes « fait tomber » cette causalité. Elle est induite dans la compréhension du lecteur par la lecture du texte.

Cependant, la méthode d'annotation proposée par Dunietz et al. est la plus élaborée que nous avons consultée et nous allons nous en servir pour la suite de nos travaux, d'abord pour nous aider à identifier des phrases comme causales (notamment avec l'aide des tests linguistiques de causalité) puis pour définir l'articulation causale à l'intérieur de chaque phrase causale. A un détail près toutefois : pour Dunietz et al., le test de contrefactualité est pris comme un critère de causalité, alors que nous souhaiterions le prendre comme une forme de causalité particulière, afin d'être cohérent avec la logique propositionnelle, celle implémentée dans une ontologie. En effet, dire que si *A* survient, *B* survient, et que si *A'* survient plutôt que *A*, *B'* survient plutôt que *B*, est un raisonnement contrefactuel ; mais d'un point de vue de la logique propositionnelle classique, si *A* est un antécédent de *B*, alors on peut seulement conclure que *non-B* est un antécédent de *non-A* (relation contraposée). Le fait que la relation de contrefactualité se détache comme modalité particulière d'association dans un raisonnement, et qu'elle doit être prise en compte comme telle dans une ontologie, sera illustré au chapitre 4 par un exemple de prise de décision durant l'accident de *Deepwater Horizon*.

En résumé, il y a deux problèmes majeurs à résoudre pour pousser plus loin ces travaux :

- cerner toutes les expressions possibles de la causalité ;
- relier des expressions de causalité entre elles dans l'ontologie : créer un « cheminement causal » dans le domaine de définition, pour éclairer des séquences d'accident.

Nous allons proposer une solution théorique à ces deux problèmes :

- le premier problème est résolu par l'utilisation d'un algorithme de classification par apprentissage supervisé ;
- le second problème est résolu par le biais d'une structuration profonde des phrases causales et une méthode d'interrogation adaptée.

3.2.3 Une méthode bayésienne de détection des expressions de la causalité

Avant de présenter notre méthode de détection par apprentissage supervisé, nous souhaitons attirer l'attention du lecteur sur une méthode linguistique, la résolution de la référence, que nous supposons efficace pour améliorer grandement la performance de la méthode que nous allons proposer. Il faut retenir ici que la résolution de la référence revient à être capable, pour un humain comme pour une machine, à identifier les unités lexicales qui renvoient vers la même signification. Les pronoms sont typiquement utilisés comme marqueurs grammaticaux de coréférences aux unités qu'ils reprennent. Nous proposons en annexe une méthode pour utiliser la coréférence grammaticale et aller au-delà.

Pour revenir à notre problématique de détection, nous avons décidé d'approcher la causalité exprimée dans un texte par un algorithme de type bayésien naïf (« Théorème de Bayes, » n.d.). Il s'agit d'un algorithme d'apprentissage supervisé, ce qui signifie par essence qu'il y a au départ de la boucle une intervention humaine.

Nous souhaitions laisser la liberté de manœuvre à l'homme pour déterminer *a priori* ce qu'est l'expression de causalité dans un texte, et force est de constater que pour le moment, malgré les avancées proprement impressionnantes des algorithmes en apprentissage non-supervisé, l'humain semble être le seul en mesure de prendre en compte virtuellement toutes les expressions de causalité dans un texte.

Le classifieur bayésien naïf est l'un des algorithmes les plus fréquemment utilisés dans le domaine de l'apprentissage automatique ou *machine learning* (Domingos, 2012), notamment dans le traitement de texte pour résoudre de nombreuses questions de recherche (McCallum and Nigam, 1998). C'est un algorithme d'apprentissage supervisé qui, à partir d'un jeu de données que l'humain aura classifié selon son besoin, va être capable de déterminer la classe à laquelle appartient une donnée issue d'un jeu nouveau. Une explication simple du code est donnée sur la page du créateur de l'algorithme, Ken Williams (« Algorithm:: NaiveBayes – search.cpan.org, » n.d.).

Il existe à l'heure actuelle de nombreuses implémentations informatiques de classifieurs bayésiens disponibles en *open source* sur l'Internet[185].

La partie naïve de ce classifieur tient au fait qu'il est fait l'hypothèse que toutes les probabilités d'occurrence d'un « mot » par rapport à un autre se valent ; or nous savons que cela est faux et qu'il existe des associations de mots beaucoup plus fréquentes que d'autres. Néanmoins, des résultats très satisfaisants sont obtenus (Domingos and Pazzani, 1997).

Nous proposons d'utiliser deux classifieurs bayésiens naïfs : le premier travaillera sur les données sémantiques, le second sur les données syntaxiques. Ils seront mis en réseau pour la résolution de la détection de phrases causales.

Dans la suite, nous présentons point par point la méthode pour la construction du modèle sémantique. Puis nous présenterons uniquement les données et les résultats

185 Source : https://github.com/search?utf8=%E2 %9C%93&q=naive+bayes+classifier&type.

attendus pour le modèle syntaxique dont le développement effectif demande des compétences informatiques que nous n'avons pas pour le moment. Eu égard au principe de compositionnalité, les deux classifieurs ont la même importance.

Nous allons présenter chaque étape de la construction du modèle, soit :

- l'annotation du texte, phrase par phrase, en marquant les phrases comme causales ou non ; il y a donc deux classes à déterminer ;
- l'utilisation d'un algorithme bayésien naïf de classification proposée dans le logiciel *KHcoder*[186] pour construire un modèle, la phase d'apprentissage ;
- l'évaluation du modèle à l'aide de la méthode d'évaluation croisée en utilisant le texte annoté manuellement ;
- l'utilisation du modèle sur un texte à classifier et découvrir de nouvelles phrases exprimant la causalité.

3.2.3.1 L'annotation du texte, la classification manuelle

Nous avons complètement annoté le texte *staff working paper* n°6[187]. Il y a cinq-cent-cinquante-trois phrases que nous avons donc classifiées en causale (C) ou non-causale (N).

Pour ce faire, nous nous sommes basés sur la méthode d'annotation de Dunietz et al. et du questionnaire de causalité de Grivaz que nous avons présentés en profondeur en amont dans ce chapitre et nous obtenons la classification suivante :

Classe	Effectif par classe
C	115
N	438

Tableau 21 : classification manuelle du document de référence

L'annotation du texte phrase par phrase est en *annexe*. Ce texte sera donc utilisé pour la phase d'apprentissage du modèle. La classification manuelle de chaque phrase a été faite ici par un seul individu (l'auteur de cette thèse). Nous avons bien conscience que cette classification initiale, qui a un rôle fondamental pour la phase d'apprentissage, est probablement peu robuste. La classification d'un texte demanderait, de la même manière que les travaux d'annotations que nous avons présentés précédemment, un cortège de classificateurs et quelques métriques pour définir l'accord inter-classificateurs à propos de l'expression de la causalité dans chaque phrase. C'est une limite à la construction du modèle (et de ses résultats) que nous allons présenter maintenant.

186 Source : (Koichi HIGUCHI, 2016, p. 61 et suivantes).
187 Source : (National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, 2011b).

3.2.3.2 La construction du modèle sémantique

Le modèle proposé ici est un modèle construit autour d'une analyse sémantique des « mots ». La donnée analysée est le lemme d'un « mot » présent dans le texte[188]. Nous allons créer pour chaque phrase une empreinte sémantique.

Avec le logiciel *KHcoder*, nous pouvons appliquer différents filtres en fonction de l'occurrence de l'unité lexicale dans le texte ou de son étiquette grammaticale[189]. Nous voyons ces réglages comme des paramètres d'entrée du modèle. Nous n'émettons aucune hypothèse sur l'utilité d'un type de morphème ou d'une unité syntaxique (une ponctuation ou une expression numérique par exemple) ni sur leur fréquence d'apparition, et c'est pour cela, que nous considérons l'ensemble des « données » accessibles et reconnues par l'algorithme de lemmatisation qui traite le texte au préalable pour le discrétiser, qualifier et quantifier les formes en présence. Nous obtenons ainsi une empreinte sémantique pour chaque phrase du texte étudié dont voici un exemple avec la phrase 12 :

> *12 « BP recognized that chief among the potential problems was the risk that methane gas escaping from the well would come into contact with sea water and form slushy hydrates, essentially clogging the cofferdam with hydrocarbon ice. »*

L'empreinte sémantique est de la sorte :

Words	POS	TF	Words	POS	TF
the	OTHER	4	from	OTHER	1
that	OTHER	2	gas	Noun	1
with	OTHER	2	hydrate	Noun	1
,	OTHER	1	hydrocarbon	Noun	1
.	OTHER	1	ice	Noun	1
BP	ProperNoun	1	into	OTHER	1
among	OTHER	1	methane	Noun	1
and	OTHER	1	potential	Adj	1
be	Verb	1	problem	Noun	1
chief	Noun	1	recognize	Verb	1
clog	Verb	1	risk	Noun	1
cofferdam	Noun	1	sea	Noun	1
come	Verb	1	slushy	Adj	1
contact	Noun	1	water	Noun	1
escape	Verb	1	well	Noun	1
essentially	Adv	1	would	OTHER	1
form	Noun	1			

Tableau 22 : empreinte sémantique de la phrase 12

188 Très précisément, nous retenons en fait l'ensemble des formes qui composent le texte pour l'analyse, c'est-à-dire tous les morphèmes, les expressions numériques quels que soient leur format, et la ponctuation, sans limite basse ou haute de fréquence d'apparition (occurrence dans le texte) pour effectuer la construction de notre modèle.

189 L'algorithme de lemmatisation et de *POS tagging* est *wsj-0-18-left3words-distsim* élaboré par le laboratoire *NLP* de Stanford (*The Stanford Natural Language Processing Group*, no date).

L'objectif est de calculer la valeur d'une probabilité conditionnelle pour chaque lemme en fonction de son appartenance à un document (une phrase) classé par un humain au préalable dans une catégorie définie. Lors de la phase d'apprentissage, le modèle va donc calculer la probabilité conditionnelle de chaque lemme (associé à son *POS tag*) par rapport à son appartenance à la classe C ou N.

> « *Dans la phase d'apprentissage, la valeur de* $\log p(w_i|C)$ *est calculée à partir de la classification manuelle.* $\log p(w_i|C)$ *reflète le score qui devrait être ajouté à la catégorie C, lorsque le mot* w_i *apparaît une fois dans le document lors de la classification automatique* » (Notre traduction Koichi HIGUCHI, 2016, p. 62).

Avec un principe de modélisation très accessible qui est :

> « *si un mot apparaît plusieurs fois dans la catégorie C dans les exemples de classement manuel [dans notre texte annoté], alors les documents [comprendre ici les phrases] qui contiennent plus d'occurrences de ce mot sont plus susceptibles d'être classés dans la catégorie C également*. » (Notre traduction, Koichi HIGUCHI, 2016, p. 62)

Enfin, la probabilité d'une phrase (un ensemble de « mots ») est :

$$p(W|C)\, p(C)$$

Elle est exprimée telle que[190] :

$$\log p(W|C)\, p(C) = \log p(w_1|C) + \log(w_2|C) + \cdots + \log p(w_n|C) \text{ (équation 11)}$$

On obtient les probabilités suivantes lors de la phase d'apprentissage du modèle (un extrait) :

Words	N	C	Variance	N (%)	C (%)
[prior probability]	9,269	7,932	0,447	53,887	46,113
the-OTHER	6,837	6,818	0,000	50,069	49,931
,-OTHER	6,404	6,142	0,017	51,042	48,958
.-OTHER	6,038	5,549	0,060	52,112	47,888
to-OTHER	5,811	5,439	0,035	51,652	48,348
of-OTHER	5,663	5,591	0,001	50,321	49,679
and-OTHER	5,521	5,238	0,020	51,319	48,681
be-Verb	5,460	5,380	0,002	50,367	49,633
a-OTHER	5,328	5,261	0,001	50,316	49,684
BP-ProperNoun	5,308	4,938	0,034	51,806	48,194
in-OTHER	5,100	4,784	0,025	51,598	48,402
have-Verb	4,913	4,855	0,001	50,293	49,707
on-OTHER	4,820	4,014	0,163	54,564	45,436
that-OTHER	4,754	4,666	0,002	50,464	49,536

Tableau 23 : extrait de la phase d'apprentissage du modèle

190 L'utilisation du logarithme permet d'éviter la production de valeurs extrêmement faibles dans le calcul des probabilités par la linéarisation des multiplications en sommes.

L'élément [*prior probability*] est un score donné à chaque phrase, peu importe les « mots » qu'elle contient ; c'est la probabilité *a priori* $\log p(C)$ définie pour chaque classe. Concrètement, c'est la valeur qui sera utilisée par la suite (dans la phase de détection) si jamais une phrase ne contenait aucune forme susceptible d'être évaluée par le modèle[191].

N, *Non-causale* et *C*, *Causale* sont donc les classes que nous avons définies pour les expressions dans les phrases. Pour chaque classe est donc calculé le résultat issu de la formule précédemment présentée. Ce résultat est appelé score du lemme sachant la classe. Le nom propre *BP* a donc un score de la classe *N* de 5,308 et de la classe *C* de 4,938. Le nom propre *BP* donne un score plus élevé lorsqu'il est classé dans la classe *N* que la classe *C*.

Pour chaque phrase analysée, le score de tous les lemmes dans cette phrase est calculé et la somme est effectuée. Le score le plus élevé de toutes les classes attribue donc la classe en question à la phrase analysée. Ce premier score donne le « poids absolu » du lemme dans l'attribution de la classe par l'algorithme.

Enfin, la variance, carré de l'écart-type, mesure la dispersion des scores de chaque catégorie entre l'ensemble des catégories. Ainsi, la variance permet de distinguer les lemmes qui affectent le plus la classification ; les lemmes les plus discriminants. La variance du nom propre *BP* a une valeur de 0,034 et est donc quasiment insignifiante ; le nom propre *BP* n'est pas particulièrement discriminant.

Pour notre modèle, nous avons effectué quelques statistiques et nous avons les résultats suivants :

Statistique	N	C	Variance	N (%)	C (%)
Nb. d'observations	2068	2068	2068	2068	2068
Minimum	0,000	0,795	0,000	0,000	21,259
Maximum	6,837	6,818	1,445	78,741	100,000
Eff. du minimum	240	1176	1	240	1
Eff. du maximum	1	1	2	1	240
1er Quartile	0,693	0,795	0,003	42,471	45,371
Médiane	0,693	0,795	0,023	46,579	53,421
3e Quartile	1,386	1,488	0,158	54,629	57,529
Moyenne	1,108	1,289	0,131	43,796	56,204
Variance (n)	0,743	0,547	0,049	329,552	329,552

Tableau 24 : statistiques descriptives du classifieur

191 Compte tenu du fait que nous prenons l'ensemble de la syntaxe en considération, nous ne rencontrerons pas ce cas dans les analyses à venir.

Avec l'illustration suivante :

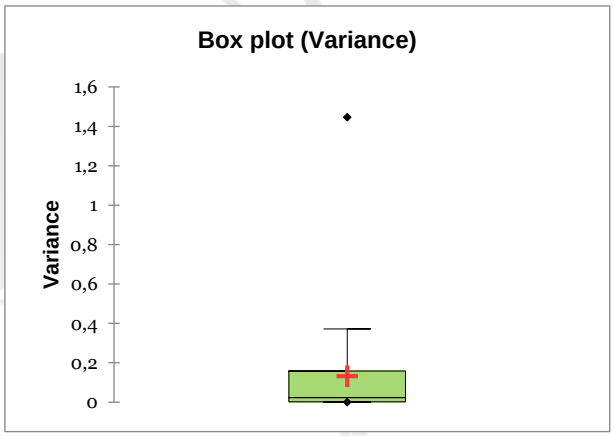


Figure 78 : boîte à moustache illustrant la variance du modèle

3.2.3.3 La validation du modèle

Nous évaluons notre modèle à l'aide de la validation croisée de type *V-fold*. Nous découpons notre ensemble de phrases en *V* dossiers avec un nombre de phrases sensiblement identiques. Chaque portion va donc être classifiée par l'intermédiaire des *V-1* restantes ; l'expérience est recommencée *V-1* fois pour que les *V* portions aient tour à tour servi à l'apprentissage et à la validation. A l'heure actuelle, le « découpage » optimal d'une population en *V-fold* est un sujet qui interpelle les statisticiens, mais qui dépasse largement les compétences de l'auteur de cette thèse. Il ne semble malheureusement pas y avoir de « recette miracle » qui fasse consensus. Nous avons choisi la valeur *V=10* pour effectuer notre validation croisée car « *[…], pour la sélection d'estimateurs, […] alors prendre V entre 5 et 10 est un très bon choix (voire optimal)* » (Arlot, 2017, p. 28).

Nous rappelons que la classification manuelle a déterminé 115 phrases causales et 438 phrases non-causales.

		learnt classified as	
human classify		C	N
	C	39	76
	N	41	397

Tableau 25 : matrice de confusion du modèle

Le modèle a quant à lui classé (39+41) 80 phrases comme étant des phrases causales (C) et (76+397) 473 comme non-causales (N).

Nous utilisons les métriques les plus communes pour la description de notre modèle que sont la précision, le rappel (ou sensibilité) et le *F-score* (moyenne harmonique ou indice de Dice) (Francis *et al.*, 1999). Pour chaque classe, on a :
La précision du modèle :

$$p_C = \frac{39}{80} = 48{,}8\% \; et \; p_N = \frac{397}{473} = 83{,}9\%$$

le rappel :

$$r_C = \frac{39}{115} = 33{,}9\% \; et \; r_N = \frac{397}{438} = 90.6\%$$

le *F-score* non pondéré ou plutôt F_1 (Sasaki, 2007) :

$$F_C = \frac{2}{\frac{1}{p_C} + \frac{1}{r_C}} = 0{,}400$$

et

$$F_N = \frac{2}{\frac{1}{p_N} + \frac{1}{r_N}} = 0{,}872$$

Une première conclusion est que le modèle est bien meilleur pour la détermination de la classe N que de la classe C.

3.2.3.3.1 La classification de nouveaux documents, la phase de détection

Nous pouvons maintenant soumettre un document nouveau, non annoté, pour que notre modèle recherche de nouvelles phrases exprimant la causalité. C'est l'objectif majeur de cet algorithme : détecter de nouvelles phrases qui expriment la causalité. Nous allons analyser deux textes issus de deux sources différentes dont le propos est le cas *Deepwater Horizon*. Nous allons nous attarder sur la classe C que nous cherchons à déterminer dans le cadre de travail de cette thèse, mais qui a pour le moment les plus faibles résultats. Nous allons évaluer seulement le critère de précision de notre modèle, à savoir, observer les résultats et marquer les faux positifs. En effet, évaluer le critère de rappel devra être fait, mais la contrainte de temps nous impose de faire un choix et nous voulons réduire le nombre de faux positifs en priorité : les phrases classées comme causales alors qu'elles ne le sont pas. L'évaluation de la causalité exprimée dans les phrases suivantes devrait se faire de la même manière que la détermination de cette dernière dans le texte de l'apprentissage.

Pour chaque texte, un fichier de classification est créé pour expliciter les résultats de la classification. Il est comparable au tableau obtenu lors de la phase d'apprentissage du modèle. Ces fichiers sont en *annexe*. Nous fournissons ici quelques exemples à titre d'illustration.

Document n°1 : https://www.britannica.com/event/Deepwater-Horizon-oil-spill-of-2010

Les phrases classées comme causales sont au nombre de six » :

1. « *Efforts in May to place a containment dome over the largest leak in the broken riser were thwarted by the buoyant action of gas hydrates -- gas molecules in an ice matrix -- formed by the reaction of natural gas and cold water.* »
2. « *Though the leak had slowed, it was estimated by a government-commissioned panel of scientists that 4,900,000 barrels of oil had already leaked into the gulf.* »
3. « *Though similar to the failed top kill, mud could be injected at much lower pressures during the static kill because of the stabilizing influence of the capping stack.* »
4. « *This entailed pumping cement through a channel -- known as a relief well -- that paralleled and eventually intersected the original well.* »

5. « *On September 17 the bottom kill maneuver was successfully executed through the first relief well.* »
6. « *A December 2013 study of living dolphins in Barataria Bay, Louisiana, found that roughly half were extremely sick; many suffered from lung and adrenal disorders known to be linked to oil exposure.* »

Voilà notre évaluation :

1. Oui ;
2. Discutable ;
3. Oui ;
4. Discutable ;
5. Non ;
6. Oui.

On a une précision dans ce cas comprise entre 50 % et 83 % en fonction des phrases « discutables ».

Document n°2 : http://www.slate.com/technology/2018/04/fire-rainbows-are-not-rainbows-and-have-nothing-to-do-with-fire.html

Les phrases classées comme causales sont au nombre de cinq :

1. « *Like most of the scenes in the movie, this one is closely based on actual events the night of April 20, 2010, when the actual Deepwater Horizon was destroyed by an uncontrolled eruption of oil and gas.* »
2. « *But as shuttle missions continued, the leaks kept getting bigger.* »
3. « *The blowout preventer was supposed to be the last-ditch defense against high-pressure gas and oil bursting out of the well.* »
4. « *BP 's Macondo Prospect blowout was a textbook case of an organizational accident.* »
5. « *The blowout caught the rest of the crew off guard as well.* »

Voilà notre évaluation :

1. Oui ;
2. Non ;
3. Non ;
4. Discutable ;
5. Discutable.

On a une précision dans ce cas entre 20 % et 60 % seulement.

3.2.3.4 Le classifieur bayésien naïf syntaxique

Nous souhaitons nous intéresser également à la syntaxe des phrases causales pour une raison fondamentale : espérer pouvoir se détacher du contexte étudié ou renforcer plus profondément les capacités d'apprentissage et donc de prédiction de notre modèle. Nous allons présenter succinctement l'idée avec un exemple.

Nous réutilisons la classification manuelle effectuée précédemment pour faire apprendre notre modèle. Cependant, pour chaque phrase, plutôt que d'appliquer une analyse sémantique et de calculer les probabilités de chaque lemme, nous proposons de calculer la probabilité conditionnelle liée à chaque relation grammaticale présente

dans chaque phrase. On reprendra les relations grammaticales *Enhanced UD* que nous avons présentées précédemment. On pourra tout de même en éliminer quelques-unes qui ne représentent pas *a priori* un intérêt pour la détection de phrases causales dans un texte (la relation *det,* les déterminants, qui n'apporte rien concernant notre problématique par exemple). Il faudra donc qualifier l'intérêt des relations grammaticales. En exemple, nous partons de la phrase suivante :

> « *BP planned to mitigate this concern once the dome had been installed by circulating warm water into the dome from the surface, so that hydrocarbons could flow up the riser unimpeded.* »

La liste des relations grammaticales de la phrase est la suivante, fournie par le *parser* de l'université de Stanford :

Relation	FROM word	TO word
nsubj	planned-2	BP-1
nsubj : xsubj	mitigate-4	BP-1
root	ROOT-0	planned-2
mark	mitigate-4	to-3
xcomp	planned-2	mitigate-4
det	concern-6	this-5
dobj	mitigate-4	concern-6
mark	installed-12	once-7
det	dome-9	the-8
nsubjpass	installed-12	dome-9
aux	installed-12	had-10
auxpass	installed-12	been-11
advcl	mitigate-4	installed-12
mark	circulating-14	by-13
advcl	installed-12	circulating-14
amod	water-16	warm-15
dobj	circulating-14	water-16
case	dome-19	into-17
det	dome-19	the-18
nmod : into	circulating-14	dome-19
case	surface-22	from-20
det	surface-22	the-21
nmod : from	dome-19	surface-22
cc	installed-12	so-24
mark	flow-28	that-25
nsubj	flow-28	hydrocarbons-26
aux	flow-28	could-27
advcl	mitigate-4	flow-28
conj : so	installed-12	flow-28
compound : prt	flow-28	up-29
det	unimpeded-32	the-30
compound	unimpeded-32	riser-31
dobj	flow-28	unimpeded-32

Tableau 26 : les relations grammaticales de la phrase d'exemple

De la même manière que nous avons créé une empreinte sémantique, nous créons une empreinte syntaxique de la phrase.

Cette empreinte syntaxique répertorie les formes en fonction de leur position dans la relation et en fonction de ladite relation ; l'empreinte syntaxique de la phrase d'exemple est en *annexe*, nous en présentons seulement un extrait :

Relations existantes entre *FROM word* (circulating-14) et *TO word*					
FROM word	*TO word*	**dobj**	**mark**	**nmod : into**	Total par relation
circulating-14	by-13		**1**		1
	dome-19			**1**	1
	water-16	**1**			1
Total (circulating-14)		**1**	**1**	**1**	3

Tableau 27 : extrait de l'empreinte syntaxique de la phrase d'exemple

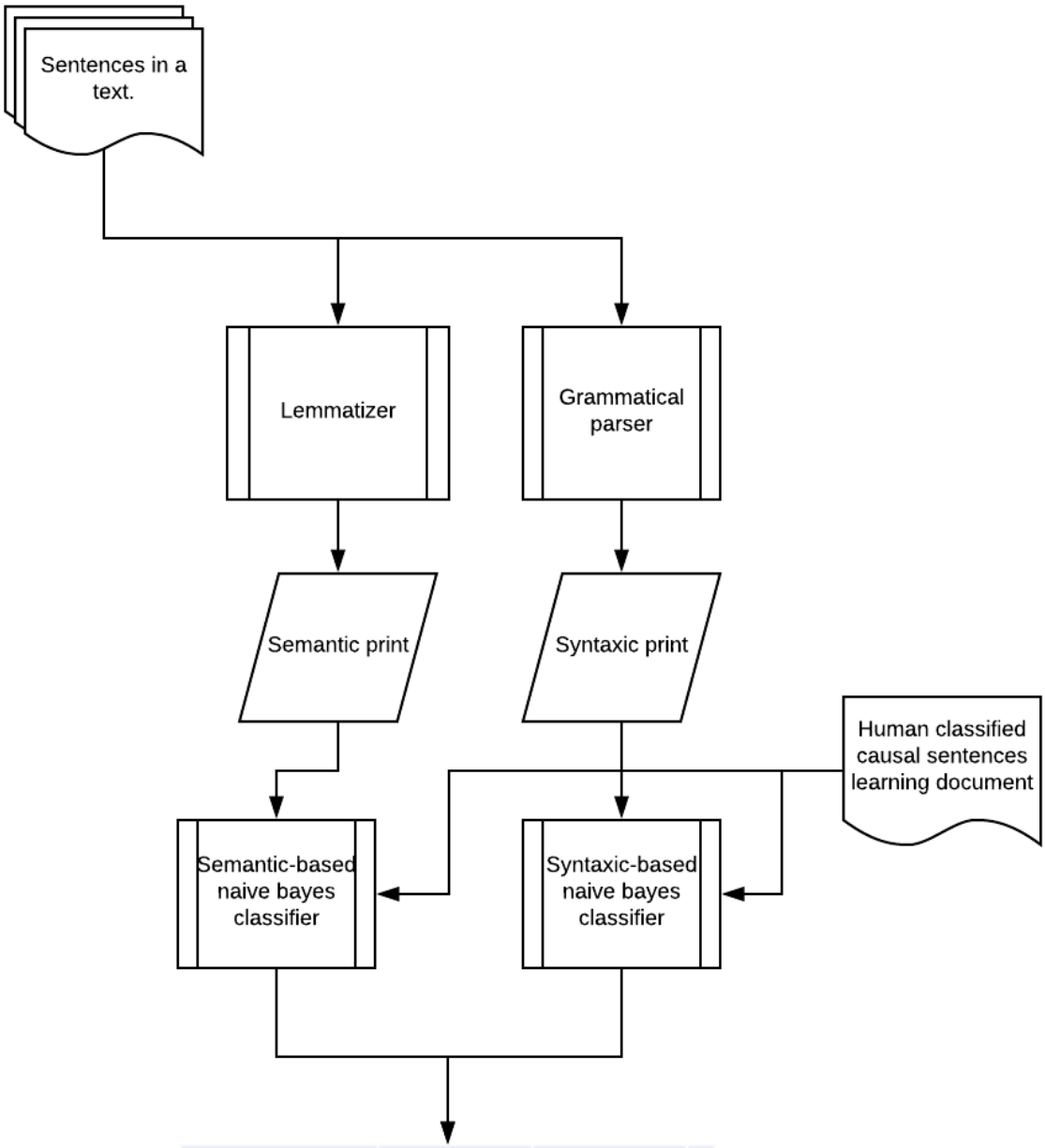
Ce tableau est un extrait de l'empreinte syntaxique telle que nous la concevons comme donnée d'entrée pour le classifieur bayésien syntaxique. Dans ce cas, pour la forme « *circulating* », en position *FROM word* (le point de départ de la relation) on a, pour chaque forme à laquelle cette dernière est reliée (donc les positions *TO word* « *water* », « *dome* », « *by* ») le type de relation et le nombre. Par exemple, « *circulating* » est relié une fois à la forme « *dome* » par la relation *nmod*.

De la même manière que pour le classifieur sémantique, à chaque type de relation seront attribuées une probabilité *a priori* et une probabilité conditionnelle en fonction de la même classification manuelle fait par l'annotateur. Chaque relation aura donc un score et les deux classes (causale et NON-causale) seront déterminées là aussi de la même manière lors de la phase d'apprentissage.

Pour la phase de validation, on procèdera aux mêmes essais et on pourra ajuster les résultats par l'évaluation de la précision et du rappel de façon identique.

3.2.3.5 La mise en réseau des classifieurs

Comme nous venons de le voir, les deux classifieurs travaillent sur les mêmes données représentées différemment. En analysant la sémantique et la syntaxe à partir d'une classification humaine établie *a priori*, nous pensons que la mise en réseau de nos deux machines peut nous permettre de détecter, avec une bonne précision et un bon rappel, des phrases de causalité. On aura donc le réseau présenté en page suivante :



semantic result	value	weight	1
Causal	53%		
NON_Causal	47%		
syntaxic result	value	weight	1
Causal	55%		
NON_Causal	45%		
final result	CAUSAL		
sentence	54%		

Figure 79 : le réseau de classifieurs bayésiens pour la détection de phrases causales

Avec seulement deux exemples, on voit que la précision varie grandement dans la classification des phrases exprimant la causalité et qu'elle est relativement faible. Nous n'avons pas évalué le rappel, mais il est peu probable qu'il se « comporte mieux ».

Nous l'avons vu, il existe de nombreux moyens d'exprimer la causalité plus ou moins fiables dans l'interprétation d'un humain et plus ou moins explicites dans la signification. C'est d'ailleurs pour cela que l'utilisation d'un classifieur bayésien qui apprend en premier lieu l'interprétation du texte que lui propose un humain nous a paru l'approche la plus appropriée, dans la limite du traitement sémantique du texte, pour détecter (en théorie) toute la richesse de l'expression de cette causalité.

La coréférence permettra également d'améliorer les résultats en réduisant sensiblement la diversité des expressions et donc en distribuant plus nettement les probabilités des expressions en question.

Enfin, un excellent moyen d'améliorer formellement ce modèle serait de lui soumettre une bien plus grande quantité de phrases causales annotées par un cortège d'annotateurs humains. On renforcerait le modèle avec des données plus robustes.

3.2.3.6 Le forage de la causalité

L'idée et l'intérêt d'aller explorer les arguments causaux dans un document écrit pour en faire émerger une élucidation des explications causales exprimées, typiquement dans un rapport d'enquête, a été montré par Travadel et al. (2018, pp. 56–70). Pour terminer, nous allons maintenant proposer une méthode de forage pour extraire les arguments de cause et d'effet qui font la nature de ces phrases et en tirer un cheminement causal. Nous présentons ici la méthode et nous n'avons pas à l'heure actuelle fixé d'implémentation informatique particulière pour réaliser le forage de causalité. Nous voyons le processus de forage en deux étapes :

1. dresser un ensemble de règles d'identification des arguments de cause et d'effet en fonction du *causal connective* ; c'est un travail de grammairien où l'on va déterminer des structures grammaticales et sémantiques utilisées pour l'expression de la causalité. Une base de données déterministe sera ainsi créée. Par exemple, la règle associée au *causal connective* « *because* » sera : « Le *chunk* portant l'argument cause est situé après le *causal connective* tandis que le *chunk* portant l'argument effet est situé avant »[192]. La phrase-type avec le *causal connective* « *because of* » sera modélisée de la sorte :

$$\{Chunk_{Forward} \xrightarrow{sequence} because \xrightarrow{sequence} Chunk_{Aft}\} \Longrightarrow^{mapping} \{[Arg_{Effect}] \xrightarrow{because\ of} [Arg_{Cause}]\}$$

A la lecture de la phrase, l'argument textuel de l'effet arrive devant le *causal connective* « *because of* » qui introduit l'argument de la cause qui arrive derrière.

192 C'est la règle de grammaire associée à l'utilisation de « *because* » comme conjonction de subordination dans une phrase pour y exprimer une relation de causalité (*Because, because of and cos, cos of - English Grammar Today - Cambridge Dictionary*, no date).

2. à la sortie du réseau bayésien, toutes les phrases causales détectées par le réseau sont passées au crible du *causal connective* identifiable et elles sont traitées en fonction. La phrase causale sera découpée selon une séquence de lecture « à la Dunietz » telle que :

$$phrase\ causale \xrightarrow{sequence} [ArgCause] \leftarrow [causal\ connective] \rightarrow [ArgEffect]$$

Deux cas se présentent : le *causal connective* est identifiable ou non. Nous abordons en détail chaque cas dans les parties suivantes.

Le *causal connective* permet donc deux choses :

- « couper » la phrase en deux *chunks,* des « morceaux » de phrases. Le *chunk* devant le *causal connective* sera appelé le *chunk forward* et le *chunk* derrière le *chunk aft* ;
- associer à chaque *chunk* un étiquetage *ArgCause* ou *ArgEffect* en fonction du *causal connective*.

Appliquons le processus de forage : nous reprenons une des phrases d'un document analysé par le classifieur, la phrase 3 du document 1 :

« *Though similar to the failed top kill, mud could be injected at much lower pressures during the static kill because of the stabilizing influence of the capping stack.* »

On pourrait illustrer la première étape de notre algorithme de la sorte :

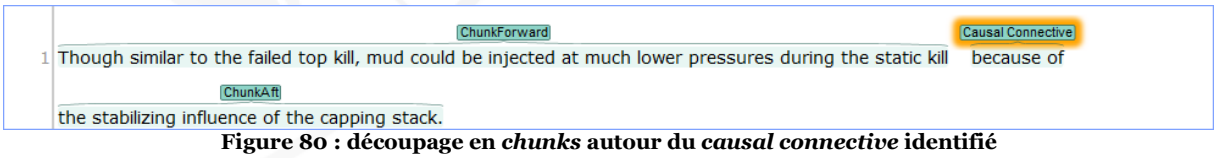


Figure 80 : découpage en *chunks* autour du *causal connective* identifié

Puis, on applique la règle liée au *causal connective*, ici identifiable comme « *because of* », ce qui donne le résultat suivant :

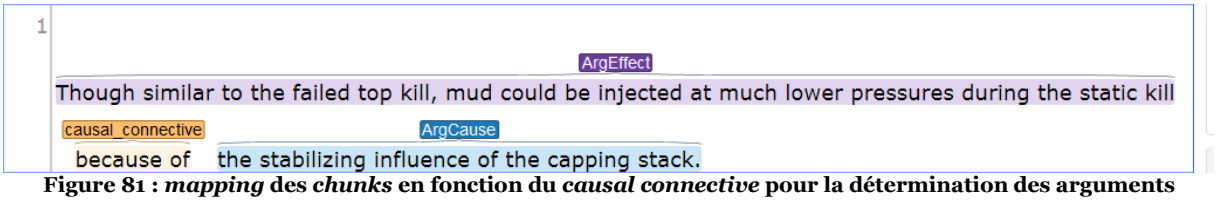


Figure 81 : *mapping* des *chunks* en fonction du *causal connective* pour la détermination des arguments

Comme on peut le constater empiriquement, il est possible que la totalité du *chunk* d'un bord ou de l'autre du *causal connective* ne soit pas indispensable à l'expression de la causalité. Aussi, nous proposons un moyen de cibler plus efficacement les expressions de causalité à l'intérieur des *chunks* en question ; nous les raffinons. Dans le *chunk*, il y a un enjeu de « raffinage » pour cibler précisément l'ensemble de mots qui serait le plus probablement l'argument cause ou effet, notamment si les phrases sont complexes ou à propositions enchâssées. Pour résoudre ce problème, nous proposons, comme pour cela, de la même manière dont nous venons d'associer un découpage structurel au *causal connective*, d'associer à chaque *causal connective* les règles de grammaire ou *a minima* le *parsing* les plus probables par lesquelles seraient associés ou structurés les arguments de cause et d'effet. On se base sur la même

grammaire universelle déployée pour l'algorithme *NER* pour effectuer l'analyse. Dans le cas de notre phrase d'exemple, on a un résultat pour le *parsing* de la sorte[193] :

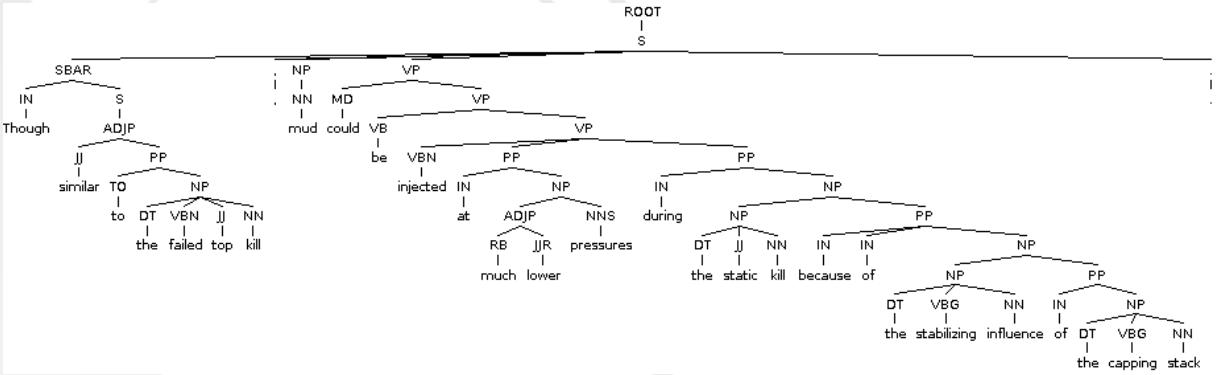


Figure 82 : *parsing* syntagmatique de la phrase d'exemple

On pourra poser comme règle pour le *causal connective* « *because of* » que les arguments cause et effet soient constitués des syntagmes nominaux les plus élevés et indispensables pour vérifier le principe de compositionnalité. Ce travail devra être l'objet d'une réflexion profonde avec des spécialistes de la langue en question pour établir un jeu de règles pertinent et efficace. Nous n'avons pas les compétences requises en linguistique, particulièrement en grammaire anglaise, pour être capable de proposer à présent cet ensemble de règles.

Le raffinage sera probablement plus long à mettre en œuvre car il faudra *a priori* définir de nombreuses règles associées au *causal connective*. C'est un apport important, mais que nous pourrons mettre en œuvre ultérieurement une fois les règles de forage établies.

Voyons maintenant le cas d'une phrase causale sans *causal connective*. Reprenons la phrase :

« Il pleut, le sol est mouillé. »

Dans le cas où il n'y pas de *causal connective* identifiable ou spécifique à l'expression de la causalité, comme dans la phrase ci-dessus, on créera pour ce type de relation un causal connective « fictif » qui pourra être tout simplement Causal_Relation. Une fois l'ensemble des phrases avec *causal connective* traité par le foreur de causalité, on procédera à la main à l'identification des *causal connectives* fictifs sur les autres phrases. On pourra associer ces nouvelles données (une sorte de nouvelle relation grammaticale) au classifieur syntaxique au fur et à mesure de la lecture humaine pour renforcer le modèle syntaxique de la phrase causale et ainsi permettre petit à petit au classifieur syntaxique de classer des phrases causales avec des *causal connective* fictifs. C'est une étape à long terme de nos travaux.

193 Obtenu avec le *parser* de Berkeley (*The Berkeley NLP Group*, no date).

On a donc finalement le *workflow* suivant pour forer la causalité :

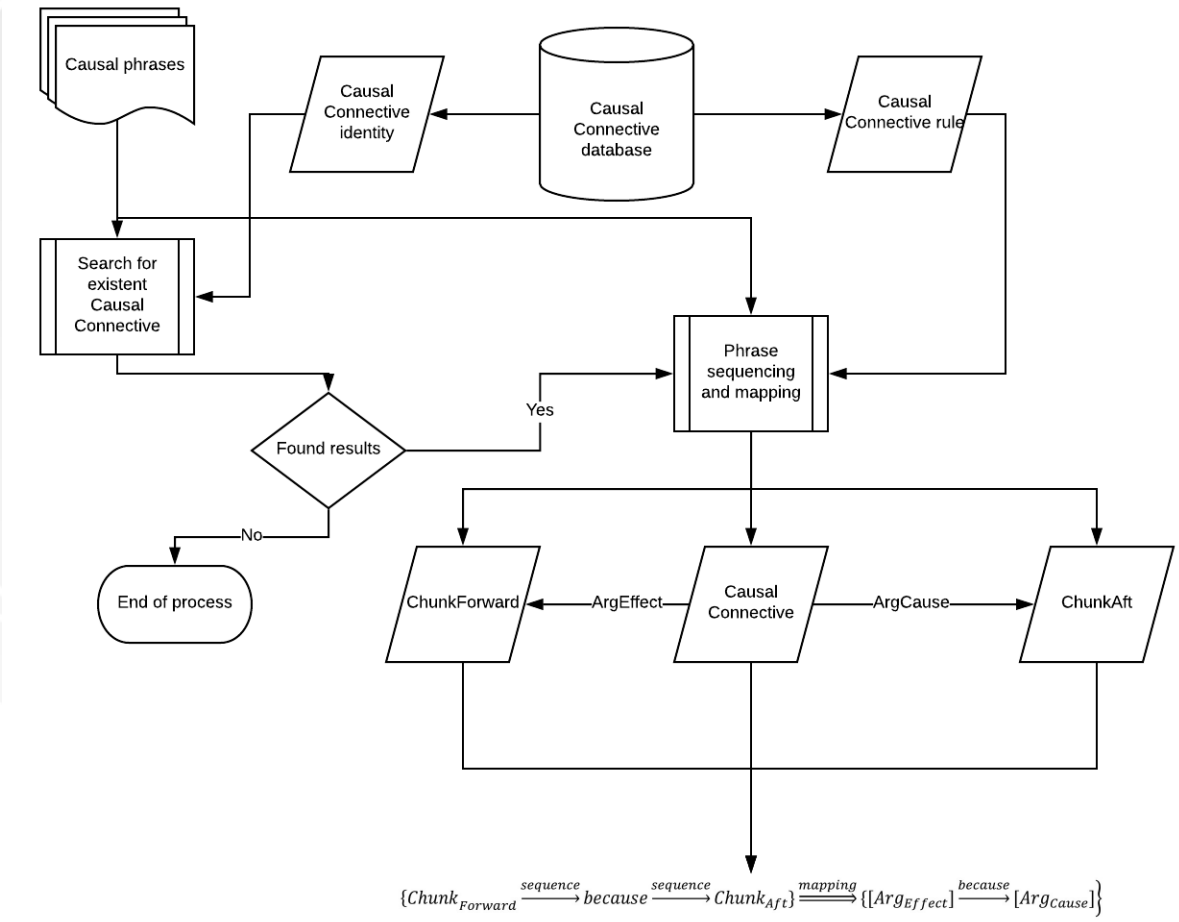


Figure 83 : *workflow* du foreur de causalité

3.3 Notre proposition : une machine qui répond à la question pourquoi ?

A ce stade, nous pensons qu'il est possible de construire une machine *question answering* qui soit capable de proposer des réponses pertinentes à la question pourquoi ? L'objectif étant de faciliter l'exploitation des connaissances sur un accident lorsqu'elles résultent d'un nombre important de données, reliées entre elles de manière complexe.

3.3.1 Aborder le cheminement causal

Nous allons maintenant rechercher d'abord la meilleure réponse possible à un questionnement causal, puis l'ensemble des relations de causes à effets exprimées dans un texte, dont nous supposons qu'il soit possible de les lier entre elles pour faire apparaître ce que nous nommons cheminement causal exprimé dans le domaine de définition.

3.3.1.1 Question-réponse

En reprenant notre exemple à propos de l'échec du *cofferdam*, on pose la question : « *Why did the cofferdam fail ?* » (1). La machine interprètera la question de la sorte : *Why* implique une recherche exclusive dans les phrases causales.

La deuxième étape va être une recherche parmi les phrases causales de celles qui ont une co-occurrence[194] des lemmes identiques avec la question posée. Ici la co-occurrence (*cofferdam*, *fail*).

En cas de question plus élaborée du type qui suit : « *Why did the hydrates form inside the cofferdam ?* » (2). On cherchera par ordre décroissant le maximum de co-occurrences les plus élevées dans les phrases causales parmi l'ensemble des combinaisons possibles de co-occurrences présentes dans la question. Dans ce cas, on cherchera l'ensemble de combinaisons possibles de deux éléments parmi les lemmes suivants : *hydrate*, *form*, *cofferdam* soit 3 possibilités : (*hydrate, form*), (*hydrate, cofferdam*) et (*cofferdam, form*).

Toutes les phrases causales seront passées au crible et le classement s'effectuera par la valeur décroissante du nombre de co-occurrences existantes dans chaque phrase. *A priori*, il n'y a pas de co-occurrence plus importante qu'une autre et elles auront toutes le même poids.

On pourra associer des filtres à l'analyse de la question pour éliminer des éléments grammaticaux ou lexicaux qui n'ont pas d'intérêt dans la recherche des co-occurrences identiques. Par exemple, la relation *UD aux* pourra être ignorée (verbes employés comme auxiliaire). Le *POS tag DET* ne sera également pas pris en compte.

Enfin, pour faciliter la tâche de l'interprétation de la question, nous utiliserons la coréférence.

Une fois les phrases causales obtenues, la réponse à la question posée est l'*ArgCause* relié par le *causal connective* de la phrase dont l'*ArgEffect* contient le maximum de co-occurrences identiques à la question posée :

$$Réponse = ArgCause_i \; tel \; que :$$

$$ArgEffect_i \ni \max[\arg(cooccurrence_{phrase_{causale}} \equiv cooccurrence_{question})]$$

Ici, on cherchera donc toutes les phrases causales qui ont dans leur *ArgEffect* les co-occurrences *(cofferdam, fail)* ou pour la deuxième question (1) ou (*hydrate, form*), (*hydrate, cofferdam*) et (*cofferdam, form*) pour la question (2). La réponse à la question sera l'*ArgCause* associé par le *causal connective* à l'*ArgEffect*.

Remarque : le premier enjeu pour notre machine est donc la reconnaissance de lemmes dans le texte de la question et dans le domaine de définition. On peut se retrouver avec les cas suivants :

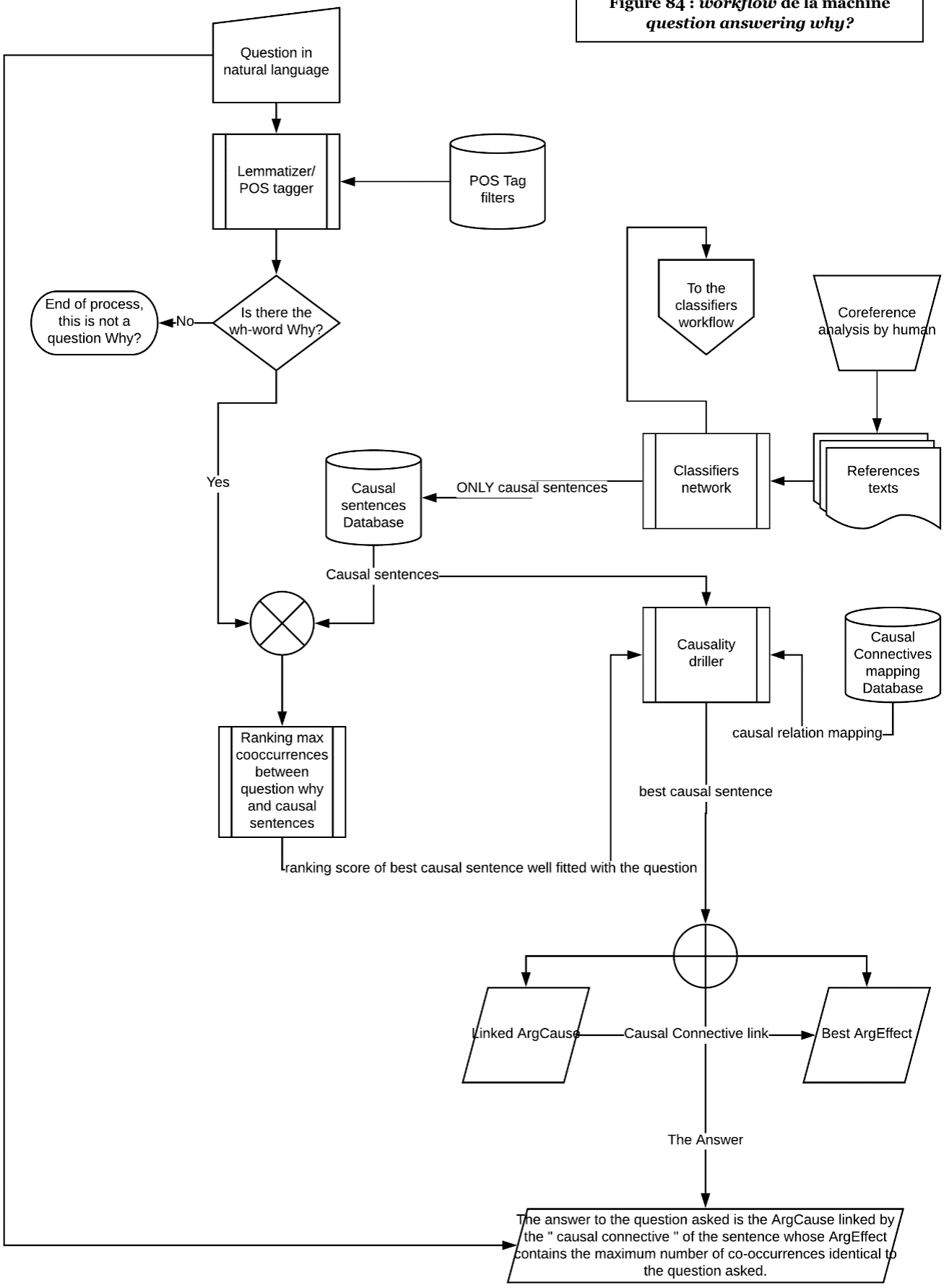
- *aucune reconnaissance de lemme dans les formes analysées ;*
- *erreur dans la reconnaissance de lemme ;*
- *reconnaissance « insuffisante » de lemme limitée par le lexique de référence.*

Pour pallier à ces problèmes, nous proposons d'ajouter un algorithme de similarité de surface en parallèle de l'analyse effectuée par l'algorithme de lemmatisation. Nous utiliserons l'algorithme de Smith-Waterman pour effectuer ce recoupement. Dans ce cas, l'algorithme comparera donc la séquence de la question posée (on supprimera seulement le *Wh Word*) avec chaque *ArgEffect* de chaque phrase causale. L'algorithme

194 Voir le chapitre 1 pour le concept de co-occurrence.

effectuera un *ranking* par valeur de similarité décroissante entre la question posée (séquence de référence) et les autres séquences. Cette analyse se fera durant la recherche des co-occurrences identiques dans les phrases causales. On prendra l'intersection des sorties de chaque algorithme (co-occurrences identiques et surface) comme résultat final.
Nous proposons à la suite un *workflow* pour l'implémentation de notre machine.

Figure 84 : *workflow* de la machine *question answering why?*



3.3.1.2 Le cheminement causal

Comme nous le disions précédemment, nous souhaitions aller plus loin dans la compréhension de l'échec du *cofferdam* et plus généralement être capable de « remonter le fil de la causalité » dans le texte étudié. On peut imaginer un algorithme, qui, à partir d'une première question posée par l'humain, va dérouler une suite de questionnements causaux en fonction des réponses obtenues jusqu'à ce qu'il ne soit plus possible de trouver de co-occurrences dans les phrases de causalité. La forme de la question pourrait être très simplement :

$$Why\ (lemme_1, lemme_2, lemme_n)?$$

On reprendra les lemmes de chaque forme de la question en appliquant les mêmes filtres que ceux de la question posée par l'humain.

Il n'y pas de structure syntaxique à proprement parler puisqu'il s'agit de rechercher des co-occurrences dans des séquences de texte déjà déterminées dans leur qualité et leur structure (sémantique et syntaxique) ; la recherche est donc grandement facilitée.

De ce fait, on construit un cheminement causal de question en question. Cela prendrait la forme d'un processus itératif qui convergerait vers l'absence de co-occurrence ou l'épuisement de tous les *ArgCause* de toutes les phrases causales pour trouver une réponse à la dernière question de la suite. En effet, une cause ne peut pas être sa propre cause, elle n'est pas réflexive.

Du coup, cela pourrait prendre la forme d'un graphique de causalité qui pourrait ressembler à cela :

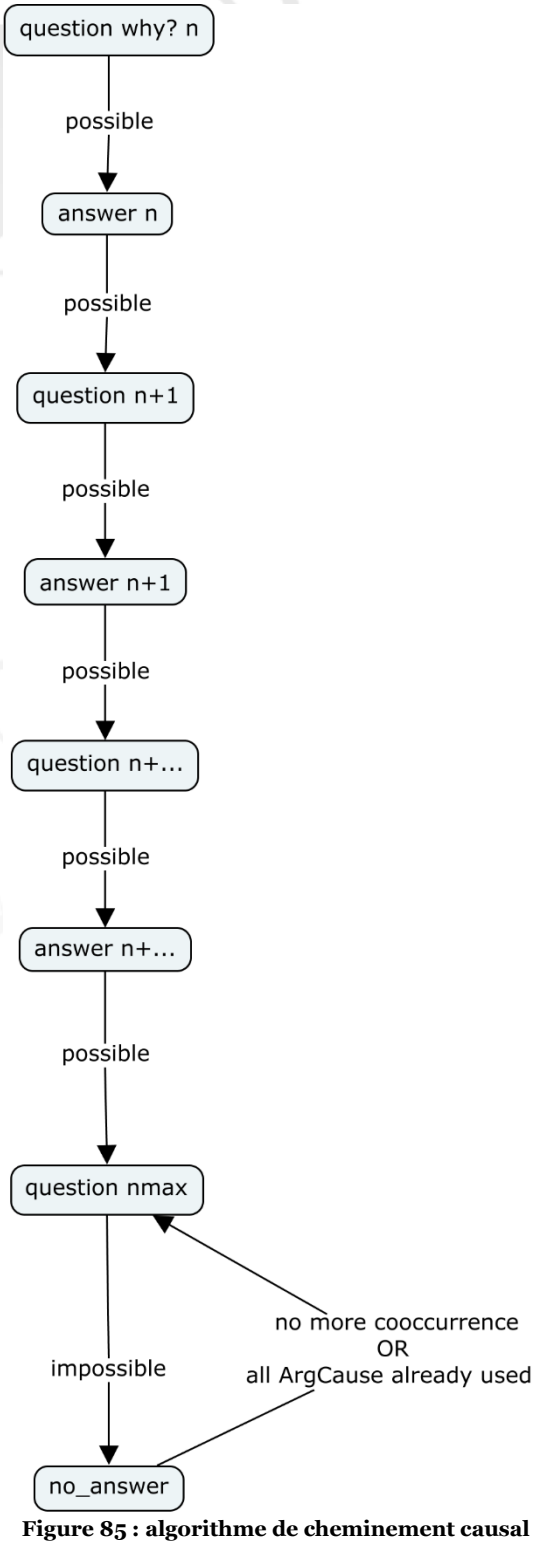


Figure 85 : algorithme de cheminement causal

3.3.1.3 Intégration de l'évènement

Nous venons de créer un cheminement qui explicite la causalité. Mais nous devons rattacher ce questionnement à l'événementialité. Nous allons utiliser le même « protocole de questionnement » pour déterminer la liaison entre une causalité et un évènement. Nous effectuerons une liaison sémantique à l'aide de la co-occurrence. Nous allons chercher à répondre à la question quand ? Nous supposons cette question être le pendant temporel de la question de causalité.

3.3.1.4 La méthode

On l'a vu, contrairement à la causalité qui est autrement plus complexe à cerner dans un texte, la temporalité est plus aisée à déterminer[195], surtout lorsqu'il y a des expressions alphanumériques de date ou d'horaire. La réponse attendue à cette question est donc de l'ordre d'une instance *TimeInterval* extraite d'une instance de la classe *Event* (une phrase d'évènement). Le domaine de définition doit évidemment comporter des phrases déterminées comme des phrases d'évènement (par notre algorithme *NER*) pour pouvoir fournir ce type d'instance.

Nous allons répondre à la question : « *When did the cofferdam fail ?* ». Pour répondre à cette question, nous proposons deux options :

- il existe au moins une phrase d'évènement qui possède le maximum de co-occurrences identiques à la co-occurrence de la question posée (ici, il faudrait donc une phrase d'évènement avec la co-occurrence *cofferdam – fail*) ;
- il n'existe pas de phrase d'évènement avec la co-occurrence cible.

a) Pour le cas *a*, la réponse proposée sera donc l'instance *TimeInterval* de la phrase en question. Dans le cas où il y aurait plusieurs phrases avec cette co-occurrence, toutes les instances seront proposées, en admettant notre méthode parfaite, cela permettra aussi d'effectuer un recoupement des informations entre différents documents par exemple. Si jamais il y a plusieurs *TimeInterval* (non-identiques) dans une phrase d'évènement, on propose que le *TimeInterval* choisi soit celui associé par une relation grammaticale (de type *nmod:…*) à l'occurrence du verbe, ici *fail*.
b) Pour le cas *b*, la recherche de réponse va passer par le cheminement causal que nous avons théorisé plus haut. En fait nous allons chercher à signifier le contenu de la question posée (« *When?* ») en rapport avec le contexte établi (le domaine de définition dans lequel le cheminement causal a été construit). Pour pouvoir répondre à la question « quand le *cofferdam* a-t-il échoué ? », il faut donc chercher à comprendre ce que la co-occurrence (« *cofferdam* », « échoué ») signifie dans notre domaine de définition. Nous allons chercher dans le cheminement causal déjà établi un lien sémantique entre la question posée (la question « *When?* ») et une phrase d'évènement. S'il existe au moins une co-occurrence identique entre une réponse dans le cheminement causal avec une phrase d'évènement, alors nous

195 La temporalité exprimée a une existence grammaticale formelle qui est le complément circonstanciel de temps.

pouvons lier cette réponse de causalité avec la phrase d'évènement et par conséquent, nous pouvons répondre à la question « *When?* ». Le cheminement causal sert d'intermédiaire de compréhension pour pouvoir répondre à la question « *When?* ».

3.3.1.5 Implémentation algorithmique

On pourra implémenter ce questionnement comme une requête implicite directement liée au questionnement causal. La machine proposera le cheminement causal et, le cas échéant, une date correspondant à l'évènement associé. Voici en page suivante une idée du graphe de causalité ainsi structuré :

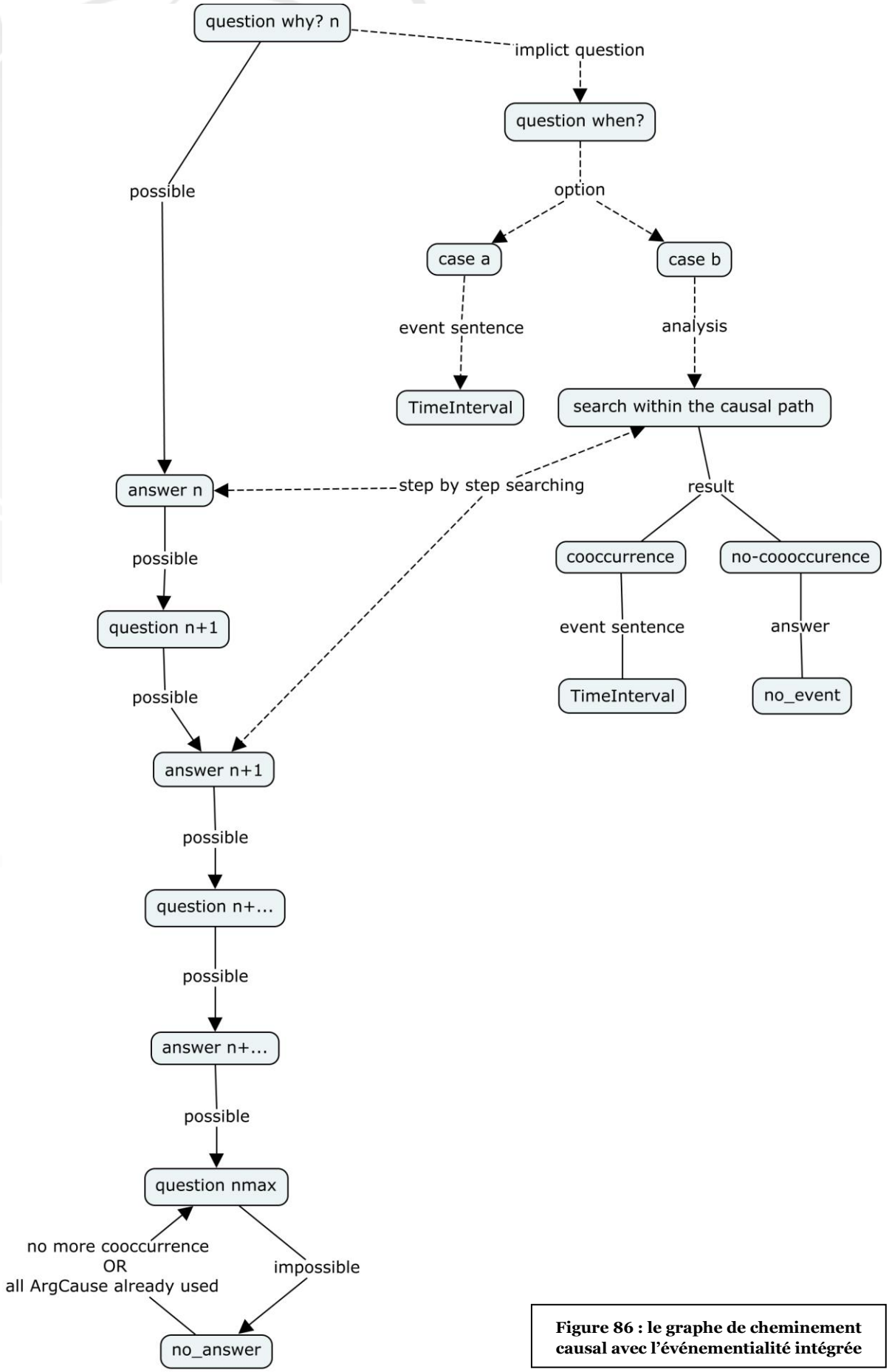


Figure 86 : le graphe de cheminement causal avec l'événementialité intégrée

3.3.2 La preuve de concept de la machine

Nous allons procéder à une preuve de concept de notre théorie et de notre méthode en l'appliquant sur le cas du *cofferdam* que nous avons déjà bien présenté. Nous allons simuler chaque étape de la méthode pour montrer la validité de notre méthode et l'intérêt des résultats obtenus. Nous voulons comprendre pourquoi le *cofferdam* a échoué et quand cela s'est passé.

Notre domaine de définition est le texte B Cofferdam de trente-huit phrases, extrait du *swp* n°6 qui relate l'histoire du *cofferdam*. Ce texte sera donc le texte de référence pour notre machine. Elle doit pouvoir y trouver des réponses pertinentes.

3.3.2.1 La question « Why » ?

Nous cherchons à répondre à la question : « *Why did the cofferdam fail ?* », et nous interrogeons la machine. Ici, nous simulons la détection de phrases causales assurée par notre réseau de classifieurs bayésiens et le forage des arguments de cause et d'effet effectué par le foreur de causalité : on obtient, par ordre d'apparition dans le texte (et qui représente la totalité des phrases causales du texte en question), les six phrases causales suivantes avec leurs arguments cause (*ArgCause*) et effet (*ArgEffect*) en fonction du *causal connective* (précisons qu'elles ne sont pas raffinées) :

Figure 87 : les phrases causales obtenues par le réseau de classifieurs dans le texte de référence

Pour l'analyse des co-occurrences identiques, nous allons effectuer une opération préalable de coréférence et on pourra considérer l'entité *cofferdam* comme étant représentée de façon équivalente par les expressions *cofferdam*, *dam* ou encore *(containment) dome*. Nous réécrirons donc le texte en unifiant la représentation de

l'entité *cofferdam* par l'expression *cofferdam* dans l'ensemble des phrases causales. Il n'y a pas de pronom qui reprenne l'entité *cofferdam* qu'il eût fallu changer.

La question posée (la question « *Why?* ») a comme co-occurrence les lemmes suivants : (*cofferdam, fail*). Il y a donc une seule co-occurrence à aller chercher. Par recoupement avec la similarité de surface, nous prenons la séquence suivante : *[Why] did the cofferdam fail ?* Nous reprenons la séquence « *thecofferdamfailed***.** ».

Nous recherchons maintenant les phrases causales qui contiennent le maximum de co-occurrences identiques avec la question posée. En parallèle, nous passons au crible de la similarité de surface les *ArgEffects* de ces phrases causales. Il y a deux phrases causales qui possèdent la même co-occurrence que la question posée :

7 « *BP planned to stage a second cofferdam on the sea floor in case the first cofferdam failed.* »

15 « *Others in the oil and gas industry were not so optimistic: Experts have told Commission staff that it was widely understood within the industry that the cofferdam effort was very likely to fail due to hydrate formation.* »

Les *ArgEffect* des phrases causales sont les suivants :

7 « *BP planned to stage a second cofferdam on the sea floor* »

15 « *the cofferdam effort was very likely to fail* »

L'ArgEffect de la phrase 15 contient la co-occurrence identique. L'ArgEffect 7 devrait être éliminé d'office puisqu'il ne contient pas la co-occurrence que nous recherchons mais nous le gardons pour la suite de la démonstration.

Les valeurs de similarité de surface des *ArgEffect* sont les suivantes :

ArgEffect	Smith-Waterman score
7	27
15	38

Tableau 28 : score Smith-Waterman des *ArgEffect*

L'*ArgEffect* de la phrase 15 a le meilleur score de similarité de surface. L'intersection de l'analyse des co-occurrences identiques et de la similarité de surface donne la phrase 15 comme meilleur résultat.

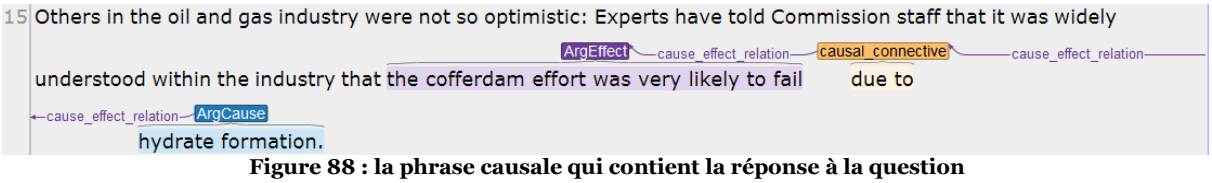


Figure 88 : la phrase causale qui contient la réponse à la question

On a donc identifié *l'ArgCause* qui porte l'élément de réponse. La réponse à la question : « *Why did the cofferdam fail* ? » est donc : « *hydrate formation* ».

Nous avons obtenu une réponse à notre question. C'est la même que l'IA d'Allen nous avait fournie. Cependant, en nous tenant au *workflow* que nous proposons pour la réalisation de notre machine, on peut penser que la réponse à cette question sera affranchie des variations qui ont trompé l'IA d'Allen.

L'enjeu de notre machine n'est pas tellement dans le système de questionnement qui est finalement assez simple à mettre en œuvre, mais bien dans la

détermination des phrases causales et de leurs arguments. Cela doit être le premier travail à fournir.

3.3.2.2 Le cheminement causal

Nous allons maintenant remonter le fil de la causalité et voir s'il existe des relations entre causes et effets qui expliquent l'argument cause identifié précédemment ; de quoi « *hydrate formation* » est-il l'effet ? Nous ne recherchons pas la similarité de surface, il faudrait chercher la séquence hydrateformation. Notre machine le fera.

Notre algorithme de cheminement causal va donc poser la question suivante : *Why (hydrate, form, formation)?*[196] On obtient quatre phrases causales :

> *12* « *BP recognized that chief among the potential problems was the risk that methane gas escaping from the well would come into contact with sea water and form slushy hydrates, essentially clogging the cofferdam with hydrocarbon ice.* »
> *15* « *Others in the oil and gas industry were not so optimistic: Experts have told Commission staff that it was widely understood within the industry that the cofferdam effort was very likely to fail due to hydrate formation.* »
> *18* « *When crews started to maneuver the cofferdam over the leak at the end of the riser on the evening of May 7, hydrates formed before the cofferdam could be put in place, clogging the opening through which oil was to be funneled.* »
> *27* « *Government officials have told Commission staff that part of the reason for the quicker-than-expected formation of hydrates in the cofferdam was the large flow volume.* »

Compte tenu que la phrase 15 ne contient qu'un seul *ArgCause* et qu'il a déjà été utilisé, nous éliminons de l'analyse la phrase en question.

Nous allons maintenant chercher dans les *ArgEffect* des trois phrases restantes ceux qui auraient cette co-occurrence. On obtient les deux phrases suivantes :

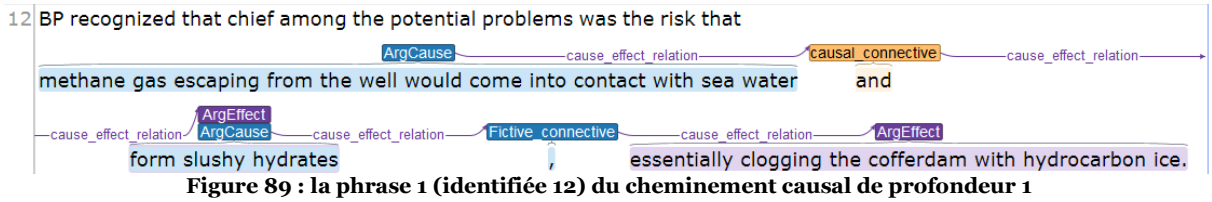


Figure 89 : la phrase 1 (identifiée 12) du cheminement causal de profondeur 1

et

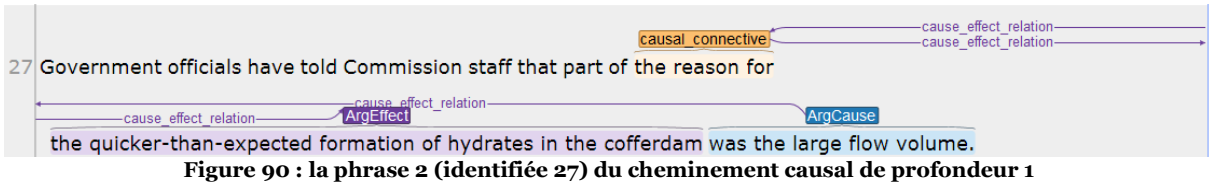


Figure 90 : la phrase 2 (identifiée 27) du cheminement causal de profondeur 1

Il y a donc deux réponses à la question : « *Why (hydrate, form, formation) ?* ». La réponse 1 : « *methane gas escaping from the well would come into contact with sea*

196 « *Form* » et « *formation* » sont considérés comme deux lemmes différents par l'algorithme de Stanford. On voit ici l'intérêt d'utiliser un algorithme de similarité de surface pour compenser ce genre de désagrément.

water », et la réponse 2 : « *was the large flow volume*». Et ainsi de suite jusqu'à épuisement des *ArgCause* ou de l'absence de co-occurrence identique. Nous pouvons d'ores et déjà tracer le cheminement causal :

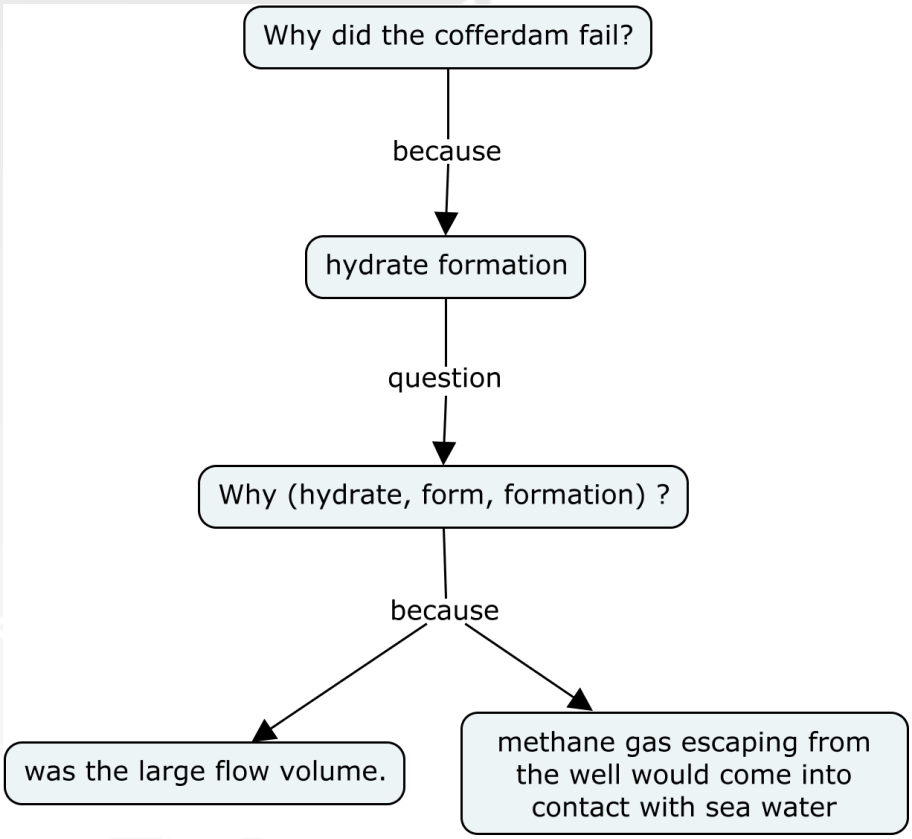


Figure 91 : le cheminement causal de l'échec du *cofferdam*

3.3.2.3 L'intégration de l'évènement

Nous cherchons maintenant à connaître la trame évènementielle du texte étudié et de rattacher si possible l'évènement à la causalité.

Nous allons maintenant simuler notre algorithme *NER* et obtenir des phrases d'évènement. Nous obtenons ces trois phrases d'évènement, les instances de la classe *TimeInterval* sont surlignées en vert :

> *1* « *On April 25, as efforts to actuate the BOP stack continued, BP began to consider placing a large containment dome, also known as a cofferdam, over the larger of the two leaks from the broken riser* ».
> *6* « *Following an MMS inspection of the Discoverer Enterprise, BP began to lower the 98-ton dome to the sea floor late in the evening of May 6*. »
> *18* « *When crews started to maneuver the cofferdam over the leak at the end of the riser on the evening of May 7, hydrates formed before the dam could be put in place, clogging the opening through which oil was to be funneled.* »

Le traitement des coréférences a été (déjà) effectué pour la représentation de l'entité *cofferdam*. Nous allons procéder de la même manière que pour les phrases causales. Nous n'effectuerons pas la similarité de surface. La question posée est donc : « *When did the cofferdam fail ?* ». Elle a comme co-occurrence les lemmes suivants : (*cofferdam, fail*). Il n'y a aucune phrase d'évènement qui partage la même co-occurrence que la question posée (cas *a*) pour nous permettre d'apporter une réponse

directe. Nous devons donc interroger le cheminement causal pour tenter d'apporter une réponse (cas *b*). La première réponse du cheminement causal est *hydrate formation*. Nous cherchons donc la co-occurrence (*hydrate, form, formation*) dans les phrases d'évènement. La phrase 18 possède cette même co-occurrence :

> *18* « *When crews started to maneuver the cofferdam over the leak at the end of the riser on the evening of May 7, hydrates formed before the dam could be put in place, clogging the opening through which oil was to be funneled.* »

Nous récupérons dans la phrase la séquence qui aura été au préalable identifiée comme une instance de la classe *TimeInterval* dans l'ontologie et nous pouvons répondre à la question : « *When did the cofferdam fail ?* », par la réponse : « *May 7* ».

A la lumière de notre connaissance du sujet, c'est une réponse juste. Nous pouvons donc construire le graphe de causalité intégré :

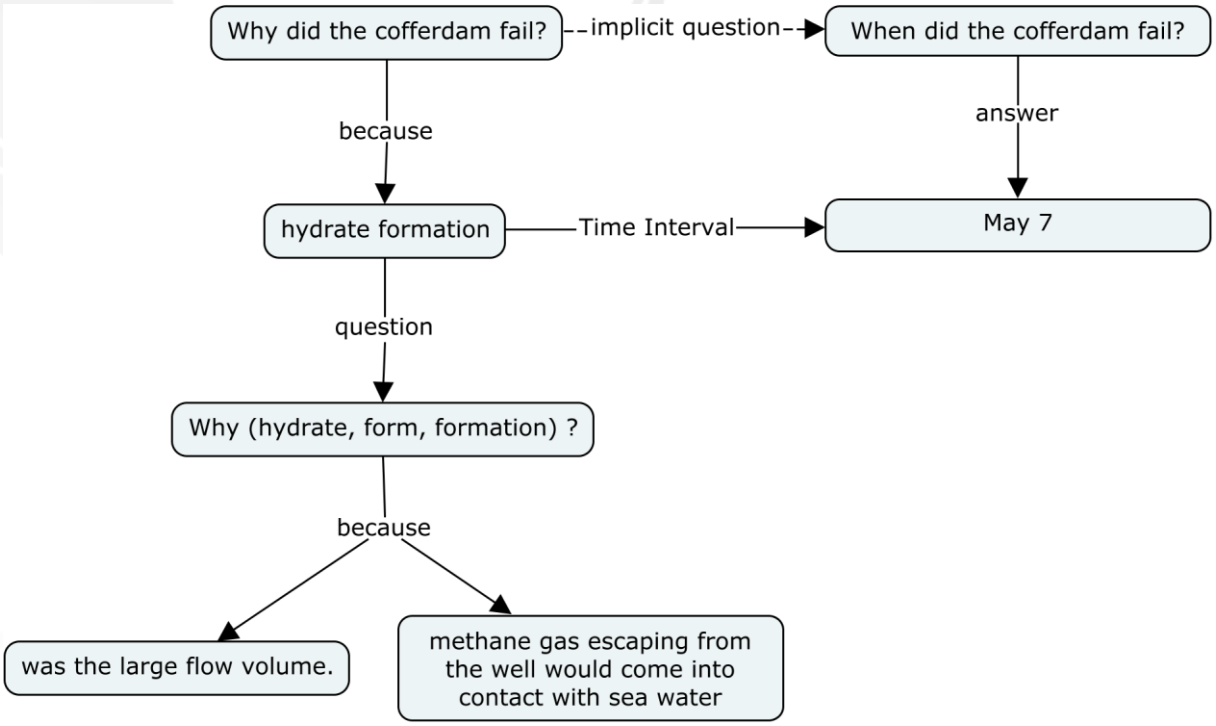


Figure 92 : graphe de causalité intégré de l'échec du *cofferdam*

3.3.3 Vers un outil opérationnel

Nous venons de montrer qu'il est possible en théorie d'atteindre les objectifs que nous nous sommes fixés. Beaucoup de concepts que nous mobilisons pour notre méthode sont déjà existants et certains sont déjà très bien implémentés à l'heure actuelle, que ce soit sous forme de code brut, d'algorithme ou de logiciel. Nous présentons ici une sorte de marche à suivre pour la mise en œuvre effective de notre méthode et la production du logiciel.

Le travail de conception et de développement peut être découpé comme suit :

- l'analyse de l'événementialité :
  - cibler les structures syntaxiques cibles propres à exprimer un évènement tel que défini dans l'ontologie *DOLCE* ;
  - constituer une base de données *Csas* ;

    - effectuer les essais d'ajustement pour l'extraction d'évènements, leur instanciation et leur mise en chronologie ;
    - implémenter par la suite la couche syntaxico-sémantique ;
- l'analyse de la causalité :
    - constituer un cortège d'annotateurs pour annoter des documents qui serviront de bases d'apprentissage pour les classifieurs ;
    - constituer un corpus de règles pour chaque *causal connective* pour implémenter le forage de causalité et la détermination des arguments. Renforcer la capacité *causal connective* fictif ;
    - faire des essais pour les réglages du « parallélisme algorithmique » entre l'algorithme de vérification de co-occurrences identiques et l'algorithme de surface ;
    - développer par la suite la capacité de raffinage du *chunk*.

En page suivante, une illustration conceptuelle de notre proposition :

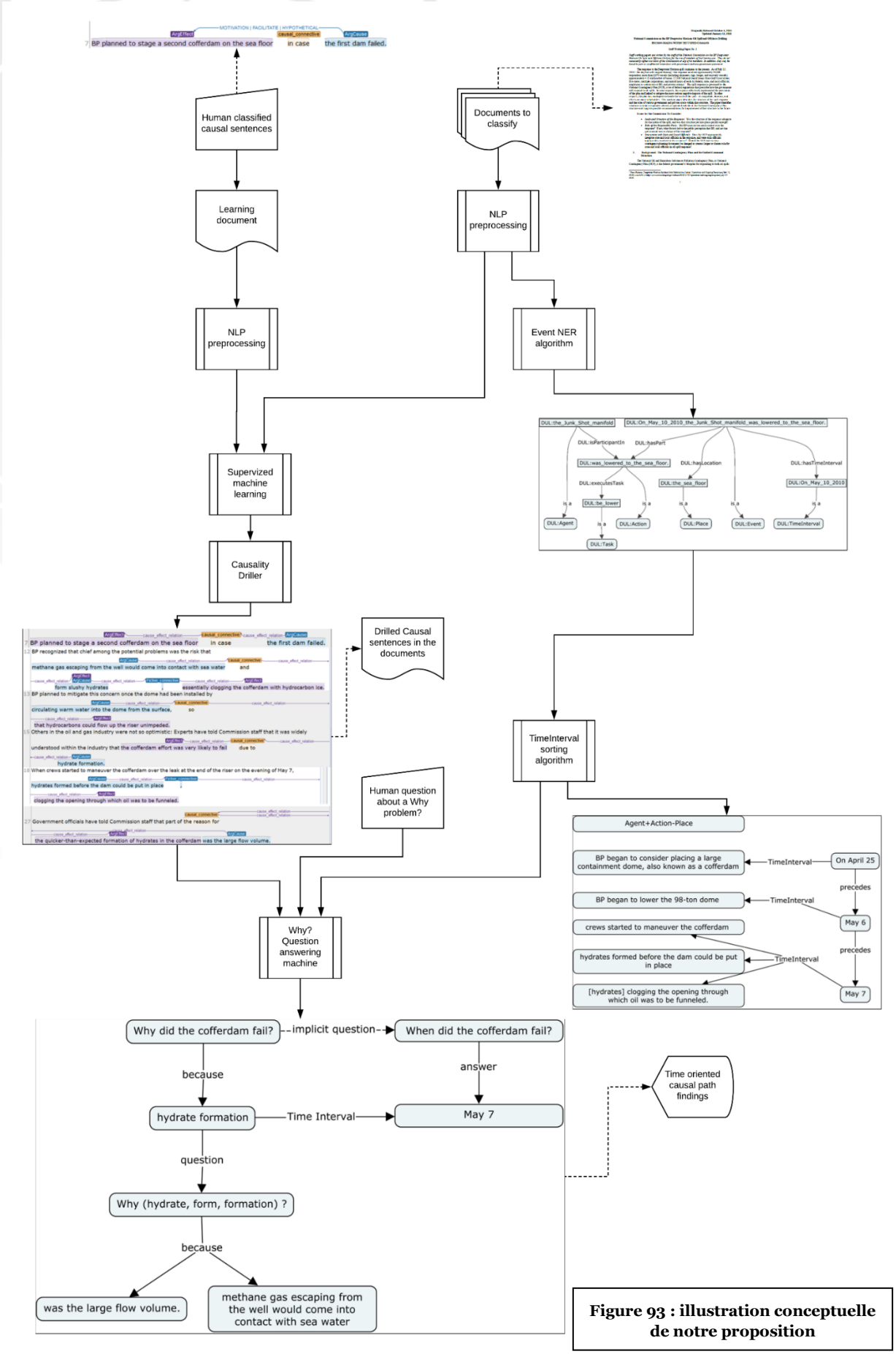


Figure 93 : illustration conceptuelle de notre proposition

3.3.3.1 A propos du raisonnement contrefactuel

La détection de la causalité contrefactuelle pourrait *a priori* poser plus de difficultés que les autres types de causalité. En effet, ses expressions pourraient être plus délicates à cerner. Cependant, en partant du principe que des annotateurs humains, formés à cerner ces expressions tout de même à la base de nombreux raisonnements d'explications causales, pourraient créer un corpus annoté spécifique de ces expressions, nous pouvons penser raisonnablement que notre réseau de classifieurs sera capable de détecter ces expressions de la même manière que les autres types de causalité « plus évidentes ». Il s'agit en fait de bien cerner lorsque l'argument ne relève pas d'une logique « classique », celle axiomatisée dans la logique propositionnelle typiquement.

3.3.3.2 Projections d'utilisations

Nous allons maintenant présenter des utilisations possibles de notre machine. En sus d'un gain de temps considérable apporté à l'utilisateur en le soustrayant à la lecture d'un document, nous pensons que les rôles de notre machine sont d'une part, la création et la capitalisation de connaissances nouvelles, et d'autre part, la mise en lumière de raisonnements de causalité amenant une plus grande profondeur de compréhension.

3.3.3.2.1 Peupler une base de connaissances spécialisée dans la causalité évènementielle

Notre machine pourra construire une base de connaissances de relations de causes à effets issues d'une multitude de documents. Il s'agit de créer d'innombrables « atomes de causalité » susceptibles d'être utilisés par la suite dans d'autres disciplines en ingénierie des connaissances ou plus prosaïquement par des spécialistes des *safety studies* à la recherche de *patterns* causaux pouvant expliciter par exemple un cas d'accident.

3.3.3.2.2 Evaluer le cheminement causal d'un document

Une utilisation principale de notre machine est de pouvoir rendre explicite le cheminement causal d'un rapport d'enquête eu égard à un questionnement humain. Ce faisant, un analyste pourra par exemple questionner la causalité exprimée et émettre de nouvelles hypothèses ou faire naître la controverse. Nous l'avons évoqué précédemment avec l'échec du *cofferdam*, et nous le montrerons en détail au chapitre suivant avec le cas du *capping stack*, l'explication causale ne tient pas toujours *via* une rationalité objective de la compréhension de la réalité, et plus généralement, nous renvoyons le lecteur au chapitre 1 qui montre toute l'étendue des possibles en matière de ce que l'on peut considérer comme une cause à un moment du processus d'enquête.

On pourra sans difficulté épistémologique élargir le spectre d'analyse de notre machine à l'ensemble des documents écrits et il pourra être intéressant de comparer entre elles différentes analyses du même cas.

3.3.3.2.3 Comparer des cheminements causaux

Une autre manière de travailler avec notre machine sera de lui soumettre différents documents qui traitent du même cas et de pouvoir ainsi comparer les cheminements causaux trouvés ; on rejoint, comme nous l'avons déjà dit, l'idée de template de Dumez (2008) qui permet grandement de faciliter l'analyse.

Une application pratique qui nous vient immédiatement à l'esprit est l'analyse de l'argumentaire déployé dans différents articles de presse à propos du même « sujet ».

3.3.3.2.4 Créer une base de données pour du machine learning

De la même manière que nous sommes en train de créer un référentiel d'apprentissage pour notre machine, les résultats qu'elle produira (les atomes de causalité) pourront eux aussi servir de données d'entrée à d'autres machines. Une première réutilisation possible de ces résultats pourra être le renforcement de notre foreur de causalité. Mais nous pensons également que d'autres applications pourront émerger à l'usage.

*
* *

Nous venons de présenter notre apport de thèse, une suite d'algorithmes susceptible de résoudre, en partie, le problème de population d'une ontologie et de proposer une représentation graphique de la causalité exprimée dans un document écrit en langage naturel. Le premier algorithme est un *NER* capable, à partir d'un référentiel ontologique, de cerner les expressions d'évènements dans les textes et de les ordonner chronologiquement. Le second est un réseau de classifieurs bayésiens (algorithmes de type apprentissage automatique supervisé) appuyé par un algorithme de mesure de similarité de surface. Ce réseau est capable de détecter, de relier et d'organiser les arguments causaux entre eux pour proposer une réponse plausible à la question pourquoi ? que pourrait se poser un humain à la lecture d'un texte. Cet apport est pour le moment théorique même si nous avons toujours gardé en tête l'objectif d'un déploiement opérationnel. Les résultats que nous allons maintenant montrer au chapitre suivant devraient être du même acabit (dans une certaine mesure, mais c'est l'idée) que ce que pourrait proposer notre suite d'algorithmes.

Chapitre 4 Application au cas Deepwater Horizon

Dans le chapitre 1, nous avons d'une part, présenté un état des lieux de l'usage du mot accident et de l'étude de l'« accident » en tant qu'objet de connaissance ; d'autre part, nous avons montré la nécessité d'un regard critique face aux sources de données et à leur exploitation dans le cadre d'une démarche scientifique sur un tel objet. Cela nous a conduits à introduire le concept d'ontologie au chapitre 2, en tant que « modèle » de représentation des connaissances. Nous avons ensuite proposé une stratégie de connexion d'une ontologie type *DOLCE* à des informations produites à partir d'un accident, en s'intéressant notamment aux notions d'évènement et de causalité (chapitre 3). Dans ce dernier chapitre, nous appliquons ces réflexions au cas *Deepwater Horizon*. Nous proposons en premier lieu une ontologie de ce cas d'ingénierie en situation extrême (4.1). Pour cela, nous rappelons comment le cas *Deepwater Horizon* s'est constitué comme un cas d'ingénierie en situation extrême (4.1.1). Puis nous présentons la phase de conceptualisation de la connaissance et le *sourcing* des données à la base de la création de notre ontologie de l'accident (4.1.2). Nous présentons ensuite des utilisations possibles de cette ontologie, notamment l'intégration de la causalité et de l'événementialité avec quelques exemples (4.1.3). Dans la section suivante, nous montrons en quoi l'utilisation d'une ontologie permet de repérer des zones d'ombre qui devraient guider la conduite de travaux de recherche (4.2). Nous montrons l'utilité de la représentation ontologique pour faire apparaître des « vides explicatifs », en particulier lors de l'étude d'une période critique de l'intervention (4.2.1). Puis, à partir de cette découverte, nous partons forer la documentation à la recherche d'une explication susceptible de combler le vide (4.2.2) et enfin, nous montrons que c'est une réflexion contrefactuelle qui est à l'origine de la prise de décision la plus critique de l'intervention : garder le puits fermé et stopper la pollution (4.2.3). Enfin, nous discutons nos résultats et présentons leurs limites (4.3). Nous présentons d'abord les limites générales de notre travail sur le cas *Deepwater Horizon* (4.3.1). Puis nous présentons les limites liées aux ontologies dans la formalisation de notre cas d'accident, et plus généralement à propos du processus de population, et plus spécifiquement dans la prise en compte de la causalité contrefactuelle (4.3.2), et enfin nous concluons cette section par les limites de notre machine dans la détermination de la causalité et du discernement de la causalité contrefactuelle (4.3.3).

4.1 Une ontologie de l'accident de *Deepwater Horizon*

Nous rappelons que nous considérons *Deepwater Horizon* comme un accident de quatre-vingt-sept jours ; cette « période d'intervention » correspond au domaine d'invalidité de la propriété P qui jusque-là caractérisait le domaine d'existence considéré, celui de la « normalité » (cf. le schéma 1 d'interprétation graphique du concept d'accident introduit au chapitre 1). A la lumière des travaux de Guarnieri et Travadel (2018), nous pouvons caractériser cette intervention comme une ingénierie en situation extrême.

Avant de proposer une ontologie de l'accident, nous en proposons une brève monographie. La liste des sources utilisées est celle déjà présentée dans le chapitre 1 ou citée de façon *ad hoc*.

4.1.1 La présentation du cas *Deepwater Horizon*

Deepwater Horizon est accident de forage pétrolier *offshore*. La survenance de la pollution n'est pas la conséquence du *blowout*[197], mais bien la perte de contrôle du puits, ouvrage de génie civil. C'est donc d'abord une faillite d'ingénierie.

Koen (1985) définit la méthode d'ingénierie comme « *une stratégie qui offrirait le meilleur changement possible en utilisant les ressources disponibles dans une situation mal comprise ou incertaine.* » Cette stratégie repose sur l'emploi d'« heuristiques », ou règles empiriquement adoptées pour leur performance. Un état de l'art (*State Of The Art, SOTA*) est défini par l'ensemble des heuristiques admises par une communauté d'ingénieurs comme leur permettant d'atteindre leur objectif de performance à un moment donné (une « époque »). Par exemple, les heuristiques mises en œuvre lors de la conception du BOP ne pouvaient pas prendre en compte le phénomène de *pipe buckling effective compression*[198] découvert par l'enquête accident du *CSB*[199]. L'état de l'art des ingénieurs à l'époque de la conception de ce type de BOP n'incluait pas ce phénomène[200] et ils ne pouvaient donc pas espérer le meilleur changement possible (en l'occurrence fermer le puits) puisque cette situation leur était tout simplement inconnue.

Pour revenir à l'intervention, des solutions d'ingénierie, issus des heuristiques de 2010, ont été mises en œuvre pour tenter de causer un « meilleur changement possible ». En l'espèce, la situation à laquelle été confrontée l'intervention était la suivante : un puits d'hydrocarbures sous-marin rentré en éruption suite à une perte de contrôle et dont le confinement a été perdu. Le meilleur changement possible était la

197 Car même s'il s'agit d'un évènement redouté, il est observé régulièrement (*Loss of Well Control | Bureau of Safety and Environmental Enforcement*, no date; *SINTEF Offshore Blowout Database*, no date).

198 Phénomène de déformation plastique subie par la tige de forage lors de l'éruption et du passage du mélange boue-gaz à grande vitesse dans le BOP, qui a empêché les mâchoires du BOP de se fermer correctement.

199 Source : (US CHEMICAL SAFETY AND HAZARD INVESTIGATION BOARD, 2014, para. 3.2.3 The AMF/deadman Fails to Seal the Well: Buckled Drillpipe).

200 Source :, compilation de document antérieurs à l'accident (*RISK ASSESSMENT OF THE DEEPWATER HORIZON BLOWOUT PREVENTER (BOP) CONTROL SYSTEM April 2000 - Final Report*, 2000).

reprise de contrôle de ce puits : stopper l'éruption et confiner la fuite. Sauf que l'état de l'art des ingénieurs antipollution était sensiblement le même depuis les accidents de la plateforme Ixtoc I (1979) et du supertanker Exxon Valdez (1989)[201] tandis que celui des ingénieurs de forage avait changé de dimension. Il y avait donc un décalage inédit entre la capacité à produire des hydrocarbures et la capacité à traiter une pollution. L'Amiral Thad Allen[202], alors *National Incident Commander*, le responsable en chef de l'intervention écrira dans son rapport :

> « *The Deepwater Horizon oil spill is the largest and most complex our nation has ever confronted, more analogous to the challenges posed by Apollo 13 than the Exxon Valdez spill of 1989.* » *(2010)*

Cependant, il existait tout de même une solution planifiée pour faire face à un puits en éruption, qui est de forer un autre puits, dit de secours, pour tenter d'intercepter la base du puits endommagé et de le sceller. C'est depuis longtemps une pratique acceptée par l'industrie et l'administration et seule la capacité financière à entreprendre ce genre de travaux a été demandée à BP avant de commencer à forer sur Macondo[203]. Cette solution fut déclenchée dès les premiers jours de l'accident, avec la mobilisation de deux plateformes de forage et nécessitait une centaine de jours[204], pour une probabilité de réussite de 95 % pour un puits et de 98 % pour deux (Greenemeier, 2010).

Cette solution, déjà acceptée et très fiable, ne va pourtant pas être considérée comme suffisante. Il va être finalement décidé de ne pas attendre la fin des forages et de lancer un vaste programme d'ingénierie pour tenter de résoudre le problème au plus vite. Nous allons maintenant présenter les solutions d'ingénierie qui ont été déployées avec des réussites variées pour tenter de résoudre le problème. Depuis le *blowout* du 20 avril, il y a eu quatre axes majeurs de développement de solutions :

- les opérations de forage. Deux puits de secours sont très vite démarrés pour tenter d'intercepter le puits endommagé (à son entrée dans le réservoir géologique) afin de dévier l'effluent et d'injecter dans ce dernier du ciment pour le sceller ;
- les opérations pour tenter de tuer le puits. Il s'agit ici de rétablir l'équilibre des pressions en injectant de la boue à haut débit dans le puits pour le « mater » et y injecter par la suite du ciment pour le sceller ;
- les opérations de collecte (de dispersion et de bioremédiation). Ce sont l'ensemble des opérations mises en œuvre pour collecter l'effluent d'une part et de limiter son impact environnemental (collecte ouverte, siphonnage,

201 Ces deux accidents sont souvent cités à tort comme des prédécesseurs, mais les dates de leur survenance servent au moins de repères temporels pour les heuristiques du moment.

202 Source : (*Admiral THAD W. ALLEN > U.S. DEPARTMENT OF DEFENSE > Biography View*, no date).

203 Source : ('BP EP - 2009 - Initial Exploration Plan, Mississippi Canyon Block', 2009)

204 D'après Doug Suttles, alors *Chief Operating Officer*, *BP* (National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, 2011b, p. 5).

barrages anti-pollution, brûlage in-situ et utilisation de dispersants à la source et en surface) ;

- les opérations de fermeture du puits. L'objectif est de proposer un moyen de fermer la tête de puits par un dispositif de même nature que le BOP défaillant pour stopper l'effluent à sa source.

Dans le cadre conceptuel de l'ingénierie en situation extrême, nous nous intéressons en premier lieu aux solutions d'ingénierie dont le cycle de vie est uniquement et exclusivement inhérent au problème à traiter, au meilleur changement à apporter dans l'urgence : il est peu plausible (comme Griffin l'entend) que ces solutions eussent été développées si l'ingénierie n'avait pas eu à faire face à ce problème inédit. Nous éliminons donc de notre champ de recherche les solutions planifiées, prévues en termes de procédures, de plan d'urgence, ou conçues comme tel. En s'appuyant sur le papier de Eude et al. (2016), nous posons comme solutions d'ingénierie en situation extrême les solutions suivantes :

- le *cofferdam* ou *containment dome* ;
- le *RITT (Riser Insertion Tube Tool)* ;
- le *CDP (Containment Disposal Project)* ;
- le *Top Hat 4* ;
- le *3-ram capping stack*.

Chacune de ces solutions est décrite en détail dans le papier cité précédemment. Ici, nous allons observer avec une « vision projet » ces solutions d'ingénierie. Nous utilisons un diagramme de Gantt de manière rétrospective pour illustrer les processus d'ingénierie mis en œuvre pendant le cycle de vie des solutions : la conception, la construction, le déploiement et l'exploitation. Le diagramme de Gantt en page suivante illustre pour chaque solution le temps d'ingénierie consacrée à sa conception, construction et mise en route (*engineering works*) et le temps d'exploitation de la solution en question (les autres périodes), son « temps efficace » et l'objectif recherché. Le code couleur (rouge ou vert) illustre le résultat obtenu (échec ou succès)[205].

205 Le succès d'une solution est donné par le consensus général des auteurs des documents de références.

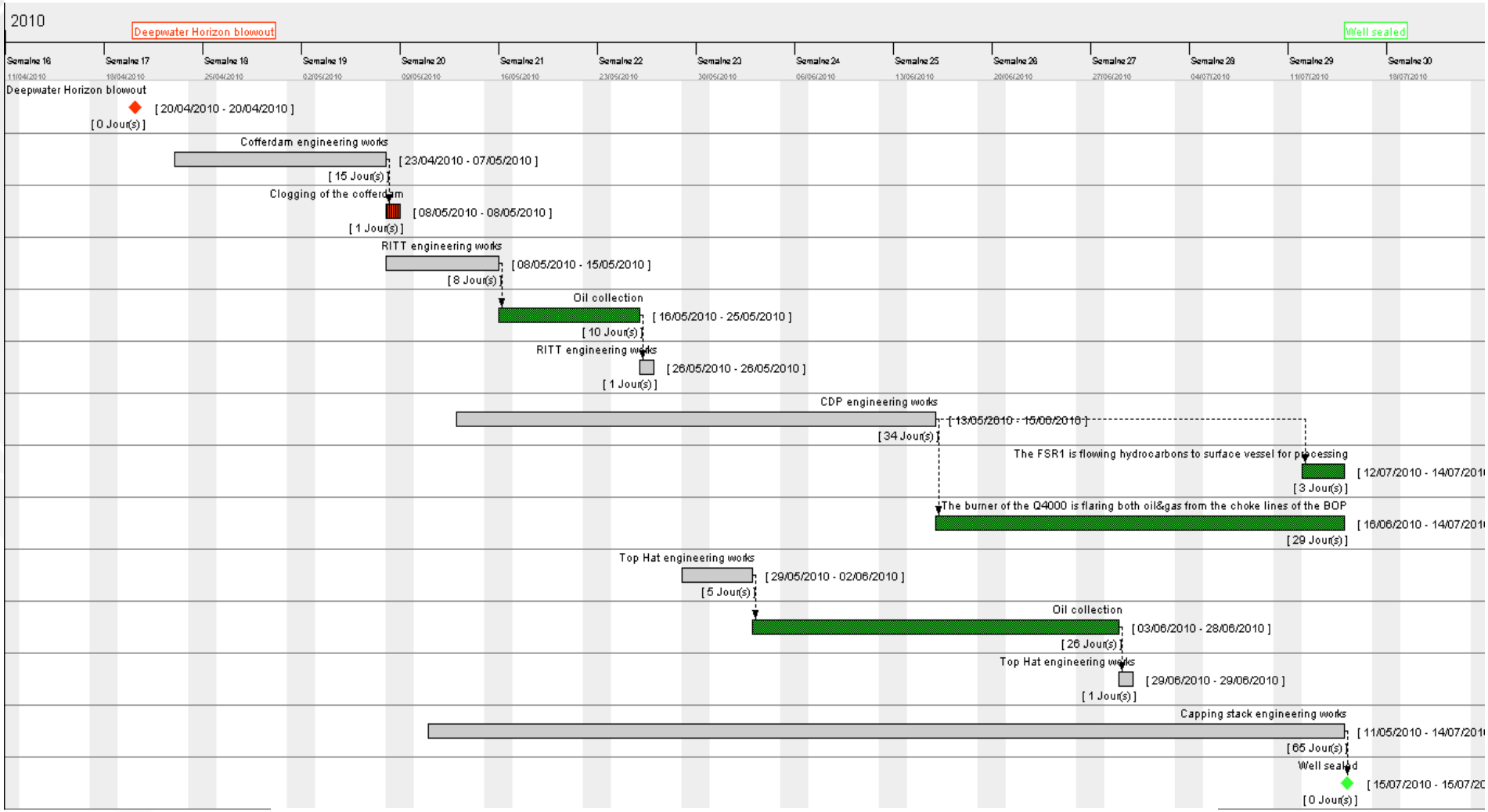


Figure 94 : diagramme de Gantt de l'ingénierie en situation extrême

Aux exceptions notables du *cofferdam* et du *top kill*, les autres solutions d'ingénierie ont été plutôt considérées comme des succès.

Les solutions d'ingénierie déployées ont permis de récupérer 810 000 bbl[206] d'huile, de fermer le puits et donc de stopper la fuite au 87ème jour de l'intervention.

En nous référant à notre définition « graphique » de l'accident (cf. chapitre 1), *Deepwater Horizon* peut être décrit par la figure 102 ci-dessous, où figurent les éléments suivants :

- la survenance de l'évènement redouté, ici, il s'agit du BOP qui ne ferme pas, la faillite de l'ingénierie « classique » ;
- la séquence des phénomènes dangereux, ici, il s'agit des fuites sous-marines ;
- la reprise de contrôle, qui implique une action humaine dans la durée ;
- l'ingénierie de l'urgence, ou l'activité d'ingénierie marquée par une pression sociétale et l'insuffisance des moyens existants.

Figure 95 : *Deepwater Horizon*, un cas de situation extrême qui dure 87 jours.

Nous allons à présent utiliser l'ontologie *DOLCE* pour formaliser la connaissance que nous avons développée sur cette activité d'ingénierie en situation extrême, en la structurant, en l'explicitant et *in fine* en apportant un enrichissement, une « méta-connaissance » à ce que nous avons déjà produit.

4.1.2 Notre ontologie de l'accident de *Deepwater Horizon*

4.1.2.1 Conceptualisation de la connaissance

Nous allons présenter la population de notre ontologie du cas *Deepwater Horizon*. Nous précisons que le processus de population a été effectué entièrement à la main[207] à l'aide du logiciel *Protégé* et en utilisant le langage *OWL*. Nous nous sommes tenus au mieux à l'engagement ontologique de *DOLCE*. Chaque instance est

206 Source : (Judge Barbier, 2015, p. 44).
207 Voir le chapitre 3 pour toute l'étendue du problème de population.

créée à partir d'un extrait du texte qui nous intéresse ou par commodité à l'aide d'une réécriture très légère. Nous prenons l'exemple suivant :

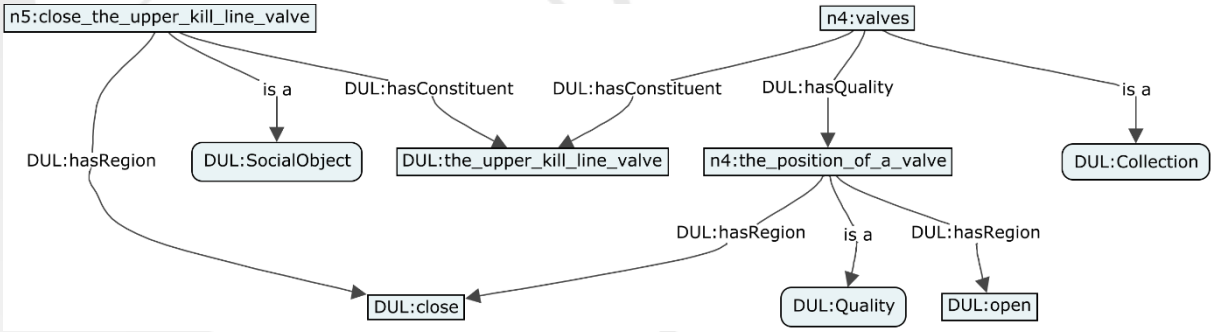


Figure 96 : structure atomique, molécules et polymères

Dans ce graphe très simplifié, nous souhaitons décrire le plus précisément possible « le fait de fermer » la vanne supérieure de la « *kill line* »[208]. Nous cherchons à créer le maximum de relations possibles par instance dans le but d'atteindre la signification irréductible du noyau atomique. Ici, en langage naturel, on doit pouvoir comprendre à la lecture de ce graphe que l'instance, un « objet social »[209], est constitué de la vanne[210] en question (instance « *the upper kill line valve* »). L'instance « *valves* » représente l'ensemble des vannes de notre monde (*Collection*) et ces vannes ont une qualité (*Quality*) qui est leur position. Cette dernière prenant deux positions possibles (des régions dans *DOLCE*), « open » ou « close ». Dans ce cas précis, pour l'instance « *close the uper kill line valve* », la position de la vanne est fermée (« *close* »). En l'occurrence, nous avons dû créer les instances « *close* » et « *open* » car elles n'existent pas en tant que telles dans le texte, mais elles sont indispensables pour enrichir la signification « générale » de notre ontologie. Pour les classes les plus abstraites (*Abstract* et *Quality* et leurs sous-classes notamment) nous avons donc créé des instances représentatives des significations contenues dans le texte (sans pour autant être exprimée de façon générale[211]) afin d'atteindre systématiquement la structure atomique. Pour effectuer ce type de graphe ontologique (et l'ensemble des graphes ontologiques, donc l'ontologie complète), nous raisonnons par analogie avec la constitution de la matière : nous créons des atomes de connaissances que nous assemblons après comme des molécules et dont certaines deviennent des polymères de grande dimension. Nous partons de l'articulation imposée par les prédicats (la logique), puis par les axiomes de classe (l'axiomatique) et enfin par les classes entre elles (l'inférence humaine ou par la machine). Cette construction se fait dans les deux sens, en fonction de l'instance que nous créons à partir du texte. C'est un travail de longue haleine et coûteux en

208 Vanne qui équipe le *capping stack* (description non montrée sur le *capping stack*).
209 C'est en fait une partie de la procédure de fermeture du *capping stack* (non montrée).
210 Nous savons que c'est une vanne car elle est un constituant de la collection « *valves* ».
211 En effet, un texte qui décrit la fermeture d'une vanne ne renvoie pas forcément au concept de vanne, de position d'une vanne et des positions possibles ; cela est souvent considéré comme « *entailed* » dans le texte et l'ontologie doit rendre explicite ce qui ne l'est pas, à condition de fournir l'effort cognitif nécessaire.

ressources cognitives. Nous avons été confrontés de plein fouet à la difficulté de la population de l'ontologie.

A l'heure actuelle, en octobre 2018, 184 instances et 376 relations ont été créées et le fichier est consultable sur demande. Nous avons donc créé une population d'instances dans *DOLCE* dont l'objectif premier est de décrire formellement les solutions d'ingénierie que nous venons de présenter. Nous souhaitons apporter à la communauté scientifique une base de connaissances renseignée, précise et exploitable, tant par l'humain que par la machine, sur ce cas d'accident. Nous allons montrer les capacités puissantes d'une ontologie et de l'intérêt de cette représentation du savoir.

4.1.2.2 Le sourcing de connaissances

Nous allons d'abord présenter un moyen de sourcer une déclaration (une assertion) dans l'ontologie. En effet, et comme nous l'avons montré dans le chapitre 1, un des enjeux majeurs de l'ingénierie des connaissances est la capacité à traiter la masse de données disponible. Aussi, le *sourcing*, le référencement documentaire qui permet la vérification de l'affirmation ontologique (la relation) par l'utilisateur est primordial. Cela permet de garantir l'intégrité de la connaissance proposée.

Dans une ontologie écrite en *RDFS* et *a fortiori* en *OWL*, il est possible d'annoter chaque instance ou chaque relation et cette annotation peut prendre différentes formes. L'annotation peut se faire au sein même de l'ontologie (des instances voire des relations annotent d'autres instances ou d'autres relations) ou en amenant de l'information « extérieure » comme du texte. La structure de l'annotation est la même que celle des prédicats. Dans *DOLCE*, son concepteur, Aldo Gangemi, s'en est servie pour décrire en langage naturel les classes et les prédicats qu'ils proposent.
Nous montrons ici un exemple d'annotation de référencement pour une relation que nous avons déclarée :

the closing sequence of the valves of the CS	**satisfies**	the shut-in procedure consisted of a series of valve turns separated by 10min rest periods to reduce the oil discharge rate to zero in a stepwise fashion

Figure 97 : annotation dans l'ontologie

Le prédicat *satisfies* est utilisé pour relier une instance de la classe *Situation* vers une instance de la classe *Description*. Nous souhaitons amener une description à propos de la séquence de fermeture des vannes du *capping stack* (*CS*). Cette description est ici un extrait mot pour mot (une citation) d'un papier où est justement décrite la séquence de fermeture des vannes ; nous devons donc indiquer la provenance de cette citation en annotant la relation.

Dans le cas de notre processus de population de l'ontologie, nous nous sommes aussi servis de l'annotation comme d'un outil de référencement pour indiquer les documents-sources qui permettent à l'utilisateur de vérifier les assertions. Nous utilisons une forme simple d'annotation qui est le commentaire ajouté (*rdfs : comment*) directement sur la relation. Nous avons le résultat suivant :

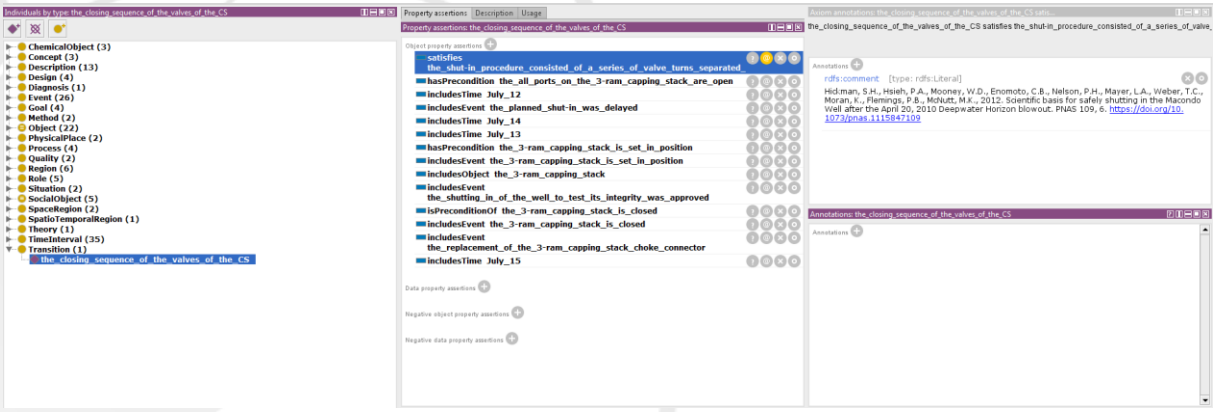


Figure 98 : une annotation de référencement pour une déclaration dans l'ontologie

Nous pensons qu'il s'agit d'une bonne pratique à systématiser lorsqu'une relation est déclarée dans l'ontologie, particulièrement si la source est rare ou difficile d'accès, ou encore si la déclaration est sujette à être vérifiée par des non-spécialistes du domaine ou prête à la controverse. En procédant de la sorte pour l'ensemble des déclarations, on rend possible une totale intégrité du processus de population.

4.1.2.3 Vers l'utilisation de l'ontologie

Nous venons de présenter la manière dont nous avons considéré notre population de l'ontologie. Au-delà de notre volonté à respecter l'engagement ontologique de *DOLCE*, nous avons axé notre réflexion sur le processus de population, à commencer par l'indispensable, mais coûteux travail de conceptualisation. Ce n'est pas tant *DOLCE* qui pose de difficulté sur ce point, mais plutôt l'effort cognitif indispensable (et imposée par l'ontologie) pour aller rechercher la structure atomique de la connaissance. Une fois le *triple*[212] créé, nous rappelons la nécessité de pouvoir explicitement identifier les sources documentaires et un moyen de pouvoir le faire dans l'ontologie.

La population d'une ontologie enrichit la connaissance extraite d'un document par le travail de conceptualisation qui oblige à une explicitation et par l'ajout de données *extra*. Cela forme une méta-connaissance du domaine que nous avions approché au départ par la lecture (naturelle) du document écrit.

4.1.3 Cas concrets d'utilisation de l'ontologie

Nous allons présenter un cas simple et un cas plus poussé de l'utilisation d'une ontologie.

4.1.3.1 La connaissance « autour » d'un thème

Nous reprenons le thème de l'échec du *cofferdam* que nous avons exploré au chapitre 3. Après un travail de recherche de mots-clés (*cofferdam*, *hydrate*) dans l'ontologie avec le logiciel Protégé, et un réarrangement pour obtenir une visualisation acceptable, voici le graphe ontologique que nous obtenons :

[212] Le *triple* est la relation ontologique irréductible constituée de trois parties : sujet, prédicat, objet.

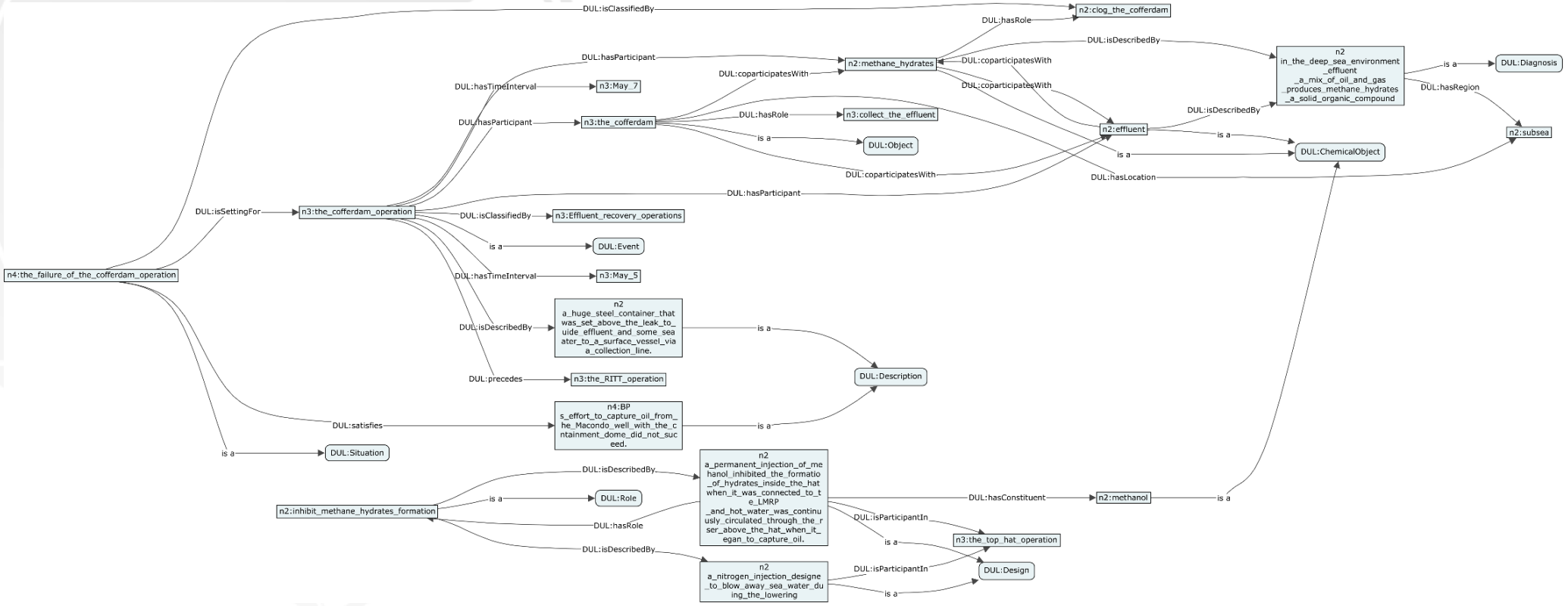


Figure 99 : l'ontologie de l'échec du *cofferdam*

On y retrouve les éléments qui ont fait l'échec de cette solution au travers du formalisme de l'ontologie. On comprend que ce sont les hydrates de méthane, dont on apprend la nature et leur mode de formation, qui sont responsables de l'échec du *cofferdam* (Classe *Situation*) en le colmatant (instance « *clog the cofferdam* » de la classe *Role*). Mais on apprend aussi qu'il est possible d'inhiber la formation des hydrates de méthane et que cette possibilité, décrite en détail (instances de la classe *Design*) sera utilisée pour la solution d'ingénierie *top hat*. Chaque liaison entre instances est définie par le prédicat correspondant pour en connaître la nature.

La création d'un « réseau de connaissances » à propos d'un sujet est une utilisation simple d'une ontologie. Le réseau de connaissances apporte au lecteur un *template* parfaitement formalisé du sujet qui l'intéresse. Il pourra être par la suite enrichi ou corrigé en fonction des découvertes à ce sujet. C'est l'utilisation la plus élémentaire, mais aussi la plus fondamentale d'une ontologie.

4.1.3.2 L'expression de la causalité dans DOLCE

Nous allons montrer maintenant comment il est possible de représenter dans une ontologie une causalité exprimée dans un texte et d'en faire un graphe de causalité. Nous rejoignons *l'ESA*[213] pour la structure de l'événementialité et la méthode d'arrangement des évènements selon la chronologie. Nous allons en premier lieu extraire des connaissances de l'ontologie par un système de requêtes. Pour cela, nous introduirons aussi ce que nous considérons, en tant qu'ontologue, comme le prédicat le plus proche de l'expression d'une causalité dans *DOLCE*. Puis, nous agencerons les résultats obtenus en utilisant le logiciel *CMAP3, Comparative Causal Mapping software*[214]*,* qui va nous permettre de dresser un cheminement causal des évènements agencés selon la causalité telle qu'elle est « codée » dans *DOLCE*. Enfin, nous présenterons certaines solutions d'ingénierie vue sous cet angle-là.

Le prédicat que nous considérons comme étant le plus à même d'exprimer le lien de causalité est le prédicat *hasPrecondition*[215]. Il est décrit de la sorte dans *DOLCE* :« *Direct precedence applied to situations. E.g., 'A precondition to declare war against a foreign country is claiming to find nuclear weapons in it'.* »
Ce prédicat subsume sous le prédicat *directlyfollows* lui-même sous *follows*. Il perd cependant la caractéristique de transitivité. Le prédicat *hasPrecondition* exprime donc une forme de relation de cause à effet entre deux évènements telle que les évènements sont agencés selon une séquence temporelle établie et une contrainte existentielle de l'évènement suivante lié à l'évènement précédent. Cela ressemble de près à une implication telle que la survenance d'un évènement *B* n'est possible si seulement l'évènement *A* existe et *a fortiori*, avec A qui précède chronologiquement *B*. On peut aussi interpréter ce prédicat comme l'expression d'un prérequis ; il faut au moins

213 Voir chapitre 3.
214 Sources (Laukkanen, 2008, 2012). Nous utilisons ce logiciel de façon très limitée, mais c'est le meilleur que nous ayons trouvé pour générer un graphe qui soit agréable, facilement exploitable et entièrement compatible avec CMAP Tools.
215 Il existe les prédicats de même nature que sont *isPreconditionOf* qui est son opposé, et *isPostconditionOf* et *hasPostcondition* qui ont la même articulation, mais qui s'appliquent dans le cas d'expressions de succession d'évènements.

l'existence de *A* pour avoir l'existence de *B*. Le *domain* et le *range* de ces prédicats sont chacun limités aux classes *Event* ou *Situation*, qui, bien qu'il ne s'agisse pas de contraintes à proprement parler, incite tout de même l'ontologue à considérer ces prédicats d'abord pour agencer une événementialité et non pas une considération de l'ordre du besoin « logistique » ; par exemple, pour faire du feu, il faut un combustible, un comburant et une énergie d'activation, la temporalité structure l'expression de la causalité dans *DOLCE*. Avec ce type de prédicat, il est possible d'exprimer des chemins de causalité comme tel :

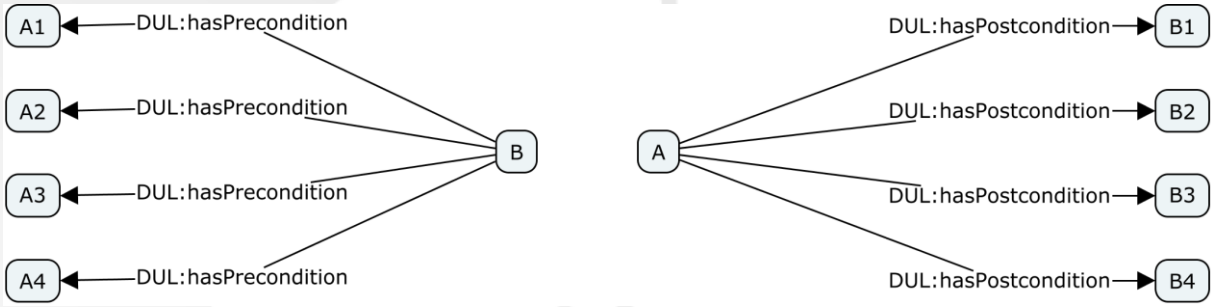


Figure 100 : la causalité dans *DOLCE*

Le cheminement de gauche est le plus proche du prérequis tandis que celui de droite se rapproche de l'implication. Evidemment, nous pouvons combiner les cheminements pour obtenir une représentation de la causalité qui pourrait être exprimée de cette manière dans un texte. On aurait par exemple le cheminement causal suivant :

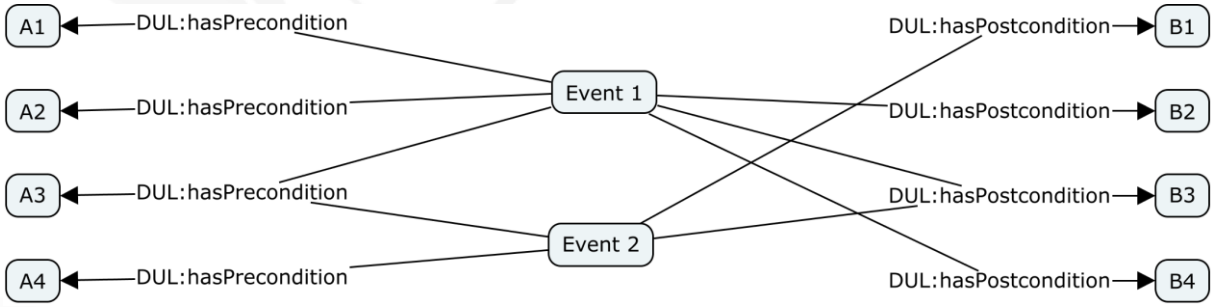


Figure 101 : un cheminement causal formalisé dans *DOLCE*

L'implicite de ce schéma reste la trame chronologique de la gauche vers la droite, nous restons bien dans un schéma de séquence tenu par la temporalité liée au concept d'évènement.

4.1.3.3 La requête de causalité

Nous allons exécuter des requêtes dans l'ontologie en utilisant l'outil *SQRWLTab*[216]. Le langage *SWRL*[217] sur lequel est basé le langage de requête *SQWRL*[218] utilise un « profil » du langage *OWL* appelé *OWL 2 RL*[219]. C'est une interprétation du langage *OWL* pour permettre à un moteur d'inférence de lire et de déduire des

216 Source : (*swrlapi: Java API for working with the SWRL rule and SQWRL query languages*, 2018).
217 Source : (*SWRL: A Semantic Web Rule Language Combining OWL and RuleML*, no date).
218 Source : (O'Connor and Das, 2009).
219 Source : (*OWL 2 Web Ontology Language Profiles (Second Edition)*, no date).

relations. Un moteur de règles comme *Drools*[220] permet de pouvoir répondre à des requêtes sur les déclarations et leurs inférences. L'état de notre ontologie est le suivant :

- *Number of SWRL rules exported to rule engine : 3,* ce sont les requêtes *SWQRL* que nous avons créées.
- *Number of OWL class declarations exported to rule engine : 76*
- *Number of OWL individual declarations exported to rule engine : 184*
- *Number of OWL object property declarations exported to rule engine : 107*
- *Number of OWL data property declarations exported to rule engine : 5*
- *Total number of OWL axioms exported to rule engine : 1456.*

Nous recherchons maintenant les instances qui auraient un lien (un prédicat) *hasPrecondition*[221] avec n'importe quelle autre instance. Comme nous avons pris le soin de bien classer les instances d'évènements et de situations, l'utilisateur qui recherchera les instances liées par le prédicat *hasPrecondition* trouvera uniquement des instances des classes *Event* ou *Situation*. La requête, très générale, est de la forme suivante :

$$Entity(?x) \cdot hasPrecondition(?x, ?w) \rightarrow sqwrl:select(?x, ?w)$$

Il y a 15 résultats :

x	w
the 3-ram capping stack is closed	the closing sequence of the valves of the CS
the 3-ram capping stack is set in position	the insertion of the transition spool
the alignment of the flange and flex joint is completed	the flex joint angle is decreased by 1 percent
the closing sequence of the valves of the CS	the all ports on the 3-ram capping stack are open
the closing sequence of the valves of the CS	the 3-ram capping stack is set in position
the failure of the cofferdam operation	the formation of methane hydrates
the insertion of the transition spool	the alignment of the flange and flex joint is completed
the planned shut-in was delayed	the leak on the choke connector of the CS
the remediation of the 2.5 degree flex joint angle	the removal of the top hat 4
the replacement of the 3-ram capping stack choke connector	the leak on the choke connector of the CS
the top hat 4 operation	the jagged cut of the riser
the top hat operation	the cut of the riser operation
the Well Integrity Test	the well is sealed

220 Source : (*Drools - Drools - Business Rules Management System (Java™, Open Source :)*, no date).
221 Que ce prédicat soit déclaré ou inféré (de même pour son équivalent opposé).

the Well Integrity Test	the shutting in of the well to test its integrity was approved
the well is sealed	the 3-ram capping stack is closed

Tableau 29 : résultats d'une requête SQWRL dans l'ontologie

Le tableau doit être lu tel que : x *has Precondition* w. Autrement dit, quels sont les prérequis w à l'existence de x pour chaque x ?

4.1.3.4 Le graphe de causalité environné de la connaissance ontologique

A partir de ce tableau de résultats, nous allons créer avec le logiciel *CMAP3* un cheminement causal où nous allons représenter ces liens sous une forme graphique. Comme nous savons que le prédicat *hasPrecondition* subsume sous *follows,* nous pouvons affirmer que l'évènement x arrive nécessairement APRES l'évènement w sur une trame chronologique. Du coup, nous pouvons donc orienter en séquence ces instances alors que nous savons que le lien représenté est plus fort ontologiquement parlant qu'un simple lien de séquence.

Avec la représentation simple $w \xrightarrow{cause\ to\ effect\ relation} x$, on obtient les graphes en page suivante :

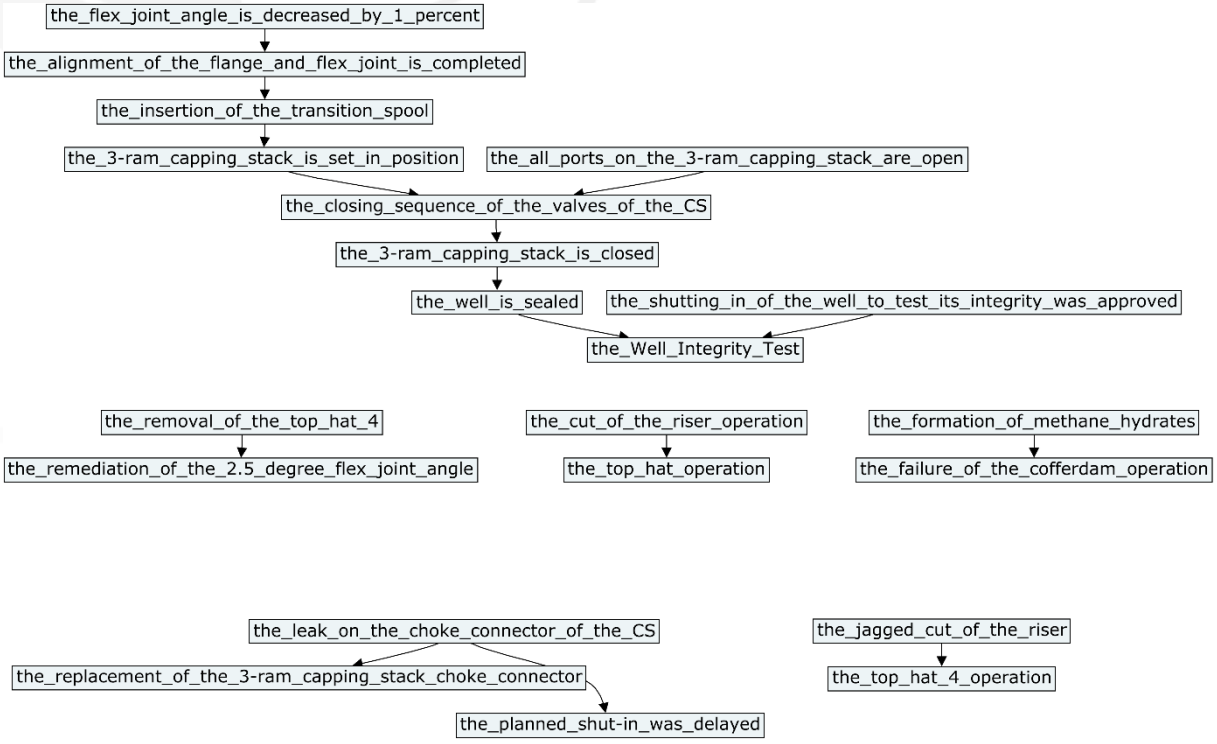
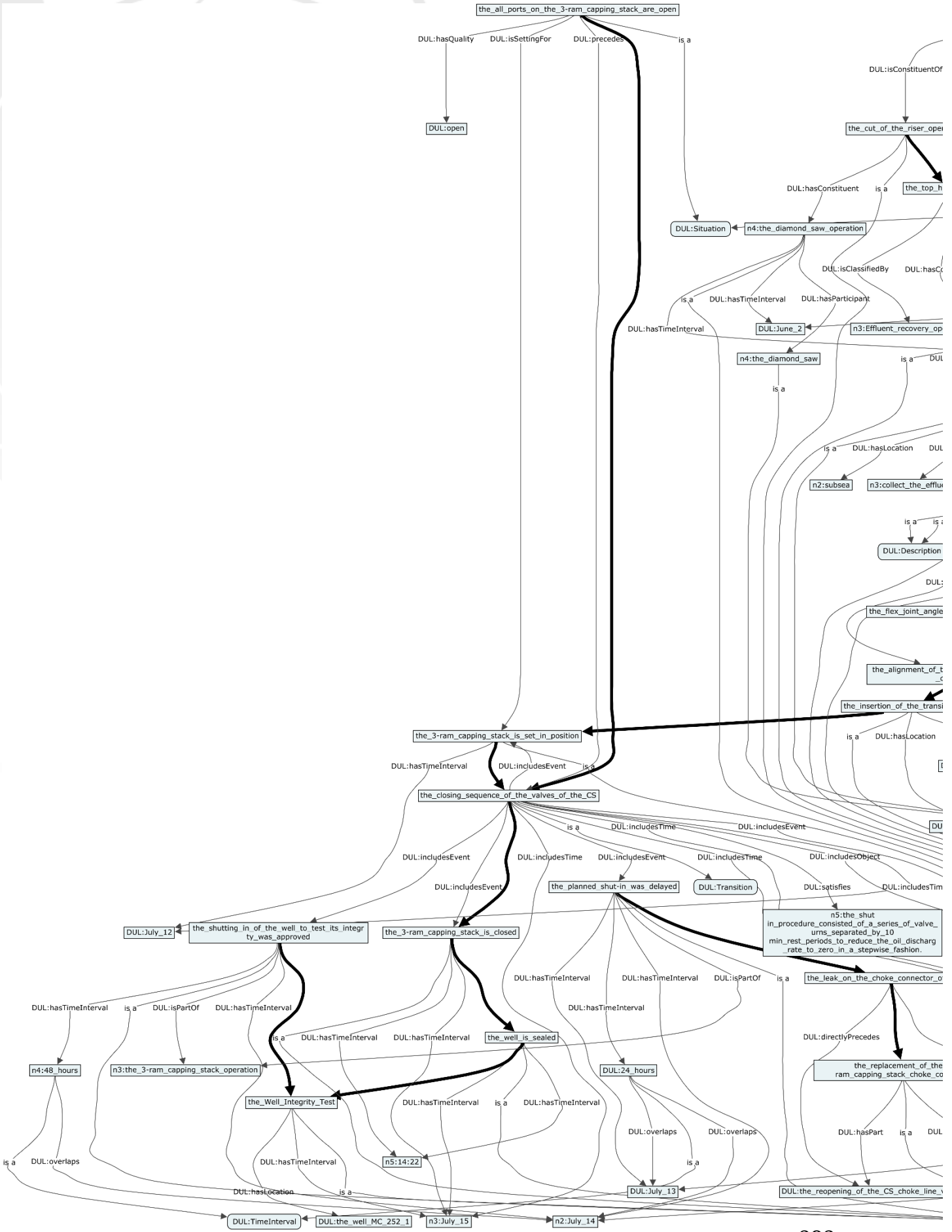


Figure 102 : les cheminements causaux créés par l'utilisation du prédicat *hasPrecondition*

L'agencement spatial des instances non reliées entre elles n'a pas de signification. On a ici dans notre ontologie six cheminements causaux exprimés. C'est un résultat intéressant car il permet de visualiser très facilement le lien de causalité entre les instances, mais on s'aperçoit assez vite, quand on connait l'ontologie (sa population) que les instances présentes ont des liens d'autre nature entre elles et avec d'autres instances non représentées. En effet, nous savons par exemple que l'instance *the remediation of the 2.5 degree flex joint angle* est reliée par le prédicat *hasPart* aux

instances *the flex joint angle is decreased by 1 percent* et *the alignment of the flange and flex joint is completed*. Nous allons maintenant associer connaissances ontologiques et expressions ontologiques de la causalité à propos du sujet étudié. La ligne noire épaisse reprend le lien de causalité montré précédemment, les autres liens correspondent à une relation ontologique entre deux instances ou entre instance et classe. Voici à la suite le graphe de causalité environné obtenu :


the_all_ports_on_the_3-ram_capping_stack_are_open
DUL:hasQuality
DUL:isSettingFor
DUL:precedes
is a
DUL:open
DUL:isConstituentOf
the_cut_of_the_riser_ope
DUL:hasConstituent
is a
the_top_h
DUL:Situation
n4:the_diamond_saw_operation
DUL:isClassifiedBy
DUL:hasC
is a
DUL:hasTimeInterval
DUL:hasParticipant
DUL:hasTimeInterval
DUL:June_2
n3:Effluent_recovery_op
n4:the_diamond_saw
is a
DUL
is a
is a
DUL:hasLocation
DUL
n2:subsea
n3:collect_the_efflu
is a
is a
DUL:Description
DUL:
the_flex_joint_angle
the_alignment_of_t
_c
the_insertion_of_the_transi
is a
DUL:hasLocation
the_3-ram_capping_stack_is_set_in_position
DUL:hasTimeInterval
DUL:includesEvent
is a
the_closing_sequence_of_the_valves_of_the_CS
is a
DUL:includesTime
DUL:includesEvent
DU
DUL:includesEvent
DUL:includesTime
DUL:includesEvent
DUL:includesTime
DUL:includesObject
DUL:includesEvent
the_planned_shut-in_was_delayed
DUL:Transition
DUL:satisfies
DUL:includesTim
n5:the_shut
in_procedure_consisted_of_a_series_of_valve_
urns_separated_by_10
min_rest_periods_to_reduce_the_oil_discharg
_rate_to_zero_in_a_stepwise_fashion.
DUL:July_12
the_shutting_in_of_the_well_to_test_its_integr
ty_was_approved
the_3-ram_capping_stack_is_closed
DUL:hasTimeInterval
DUL:hasTimeInterval
DUL:isPartOf
is a
the_leak_on_the_choke_connector_o
DUL:hasTimeInterval
is a
DUL:isPartOf
DUL:hasTimeInterval
DUL:hasTimeInterval
DUL:directlyPrecedes
is a
DUL:hasTimeInterval
DUL:hasTimeInterval
the_well_is_sealed
the_replacement_of_the
ram_capping_stack_choke_co
n4:48_hours
n3:the_3-ram_capping_stack_operation
DUL:24_hours
the_Well_Integrity_Test
DUL:hasTimeInterval
is a
DUL:hasTimeInterval
DUL:overlaps
DUL:overlaps
DUL:hasPart
is a
DUL
is a
DUL:overlaps
DUL:hasTimeInterval
n5:14:22
is a
DUL:hasLocation
is a
DUL:July_13
DUL:the_reopening_of_the_CS_choke_line_
DUL:TimeInterval
DUL:the_well_MC_252_1
n3:July_15
n2:July_14

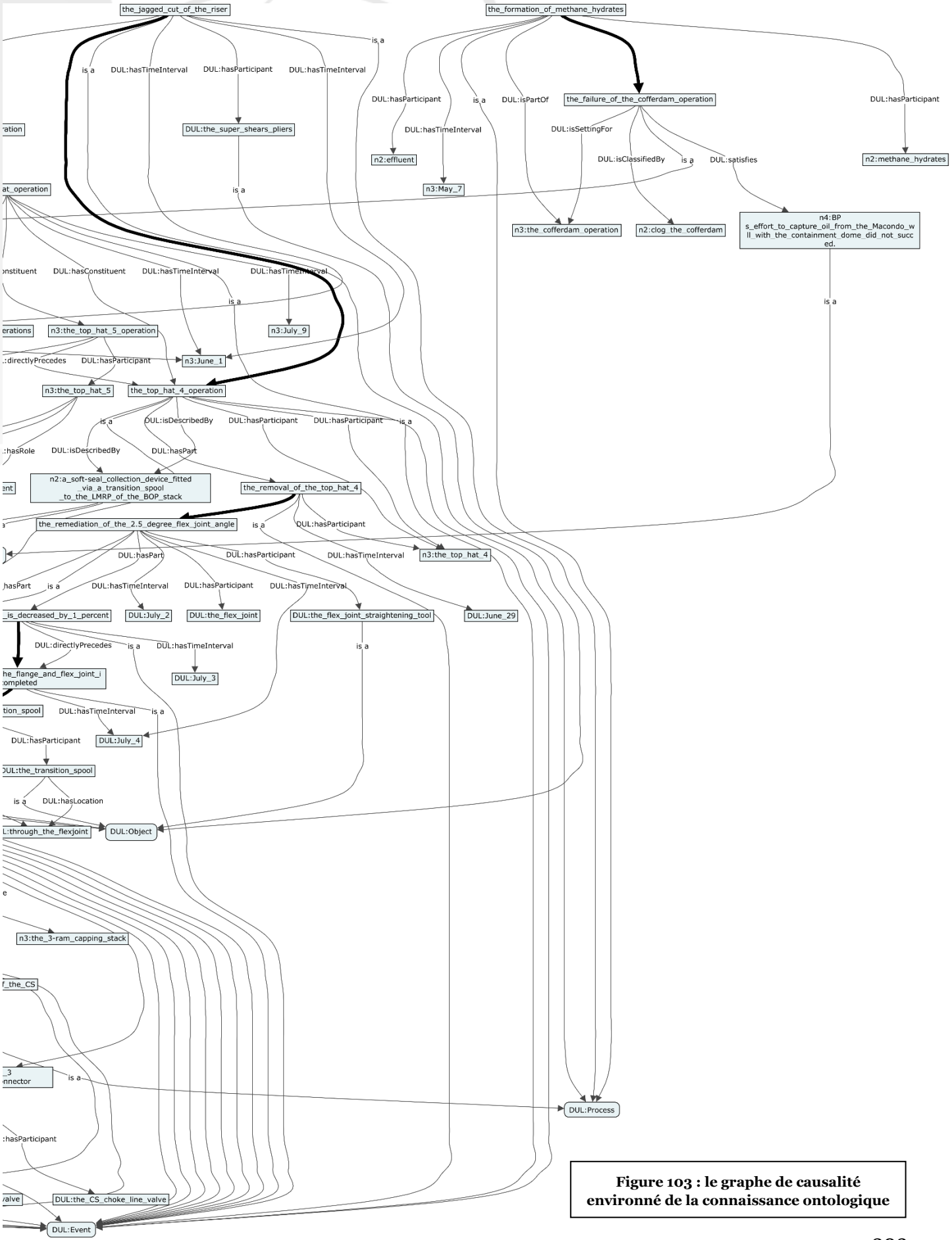


Figure 103 : le graphe de causalité environné de la connaissance ontologique

Nous venons de montrer qu'il est possible de formaliser la causalité exprimée entre différents évènements dans le but de proposer un graphe ontologique « total » où sont visibles l'ensemble des connaissances à propos d'un sujet et les expressions de causalité, sous forme de prérequis ou d'implications, qui amène un niveau de compréhension supérieur au sujet étudié. Nous allons maintenant resserrer notre travail de formalisation ontologique sur une séquence très particulière de l'intervention.

4.2 L'ontologie pour orienter la recherche

Lors de nos travaux de recherche sur ce cas, nous avons cherché à identifier « le moment » où la solution proposée face au problème s'est avérée efficace, l'instant où le problème a été résolu, où l'intervention a permis de reprendre le contrôle de la situation. Pour l'accident de *Deepwater Horizon*, on apprend, à la lecture des rapports d'enquête et d'autres sources que nous avons identifiées[222], et évidemment *a posteriori*, qu'il y a eu des jalons critiques pendant cette intervention. Ces jalons marquent un tournant dans la lutte, c'est-à-dire qu'il y a un avant et un après, que des changements bénéfiques ont été effectués sur le long chemin de la résolution. Si l'on revient au problème auquel a fait face l'intervention que nous avons explicitée plus tôt, à savoir un puits d'hydrocarbures sous-marin rentré en éruption à la suite d'une perte de contrôle et dont le confinement a été perdu, le meilleur changement possible s'est opéré le 15 juillet 2010. Le 15 juillet 2010 est une date fondamentale puisqu'il s'agit du premier jour depuis le 20 avril où la fuite sous-marine est stoppée. En effet, l'équipe d'intervention, dirigée par l'amiral Thad Allen, alors *National Incident Commander* (*NIC*), le Commandant en chef[223], et composée des ingénieurs de BP et des scientifiques du gouvernement américain (*Government Led Science Team, GLST*), sous la direction du Pr. Steven Chu, a déployé avec succès le *capping stack*, un équipement qui a la capacité de fermer le puits de façon définitive. Cette installation et la fermeture de cet équipement ne se sont pas faits sans difficulté, mais le résultat obtenu est sans commune mesure avec toutes les autres solutions d'ingénierie déployées précédemment puisqu'elle marque enfin la fin du déversement d'hydrocarbures dans l'environnement. Mais le répit aurait pu être de courte durée car un nouveau problème fit rapidement son apparition, la pression dans le réservoir ; les scientifiques et ingénieurs auraient pu avoir à rouvrir le puits dans les 24 heures. Nous appellerons cette séquence, qui s'étend du 14 au 16 juillet compris, « séquence du *capping stack* ». C'est la discussion de cette deuxième partie. En utilisant les ontologies, avec *DOLCE*, nous allons formaliser et mettre en évidence l'insuffisance des explications de cette séquence proposées dans les documents disponibles sur l'accident. La question que nous nous posons est la suivante : *pourquoi la WIT gardera elle finalement le puits fermé, alors que les quelques données objectives pouvaient inciter à le rouvrir ?*

222 Voir le chapitre 1.
223 Pour bien comprendre les rôles et les responsabilités au sein de *l'Unified Command*, lire en particulier les rapports (Allen, Thad W., 2010; United States Coast Guard, 2011).

Au cours de nos recherches, nous avons abordé intuitivement cette séquence, avec une connaissance experte du cas, ce qui nous a conduits à l'isoler comme un jalon crucial et, en même temps, peu analysé à travers les travaux scientifiques ou les investigations officielles. Dans cette section, nous voulons montrer qu'il est également possible d'aboutir à la même conclusion d'une insuffisance des explications relatives à cette séquence, mais en utilisant l'ontologie de manière appropriée.

4.2.1 Résolution graphique d'un incident d'explication dans la connaissance

Nous donnons d'abord une brève description de la situation. Le 12 juillet, l'équipe d'intervention, la *Well Integrity Team (WIT)*, composée de membres de la *GLST* et d'ingénieurs de BP, déploie au-dessus du BOP[224], une solution d'ingénierie : le *3-ram capping stack*[225], considérée comme « *la plus grande réussite de tout l'effort d'ingénierie de l'intervention* » d'après Doug Suttles, BP Chief Operating Officer of Exploration and Production. Cet équipement, une sorte de BOP, est un système *cap or seal*. Il peut donc collecter l'effluent d'une manière étanche (c'est-à-dire sans contact avec le milieu extérieur, en l'occurrence l'eau de mer) ou sceller le puits par l'intermédiaire de son jeu de trois mâchoires étanches.

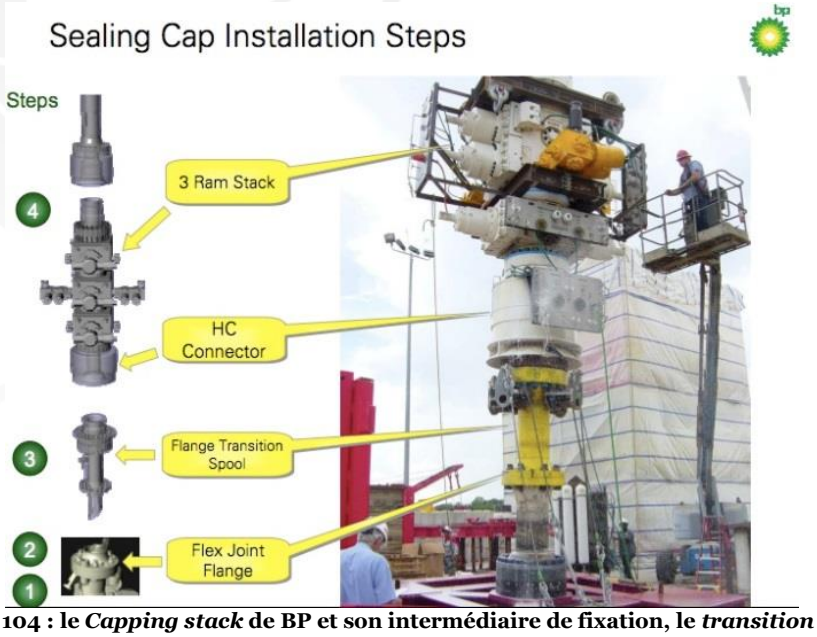


Figure 104 : le *Capping stack* de BP et son intermédiaire de fixation, le *transition spool*

Pour l'équipe d'intervention, après une période de trois mois d'efforts gigantesques et de succès très relatifs, cela signifie qu'il est enfin possible de fermer le puits et d'arrêter le déversement d'hydrocarbures, c'est le meilleur changement possible à apporter à la situation.

224 Plus exactement au-dessus du *LMRP*, le support physique étant un *transition spool* inséré dans le *LMRP* pour accueillir le *capping stack* et ainsi supporter son poids et rigidifier la connexion.

225 On voit aussi ici et là l'expression « *sealing cap* » qui correspond au même objet technique. Nous l'appellerons *capping stack* par la suite.

L'opération de fermeture du puits est déclenchée le 14 juillet 2010[226] sur ordre formel de Thad Allen. Les vannes sur le *capping stack* sont fermées selon un ordre précis et la dernière vanne est fermée par incréments.

Le capping stack est fermé le 15 juillet 2010 à 14h22[227]. Voilà le graphe ontologique de la situation :

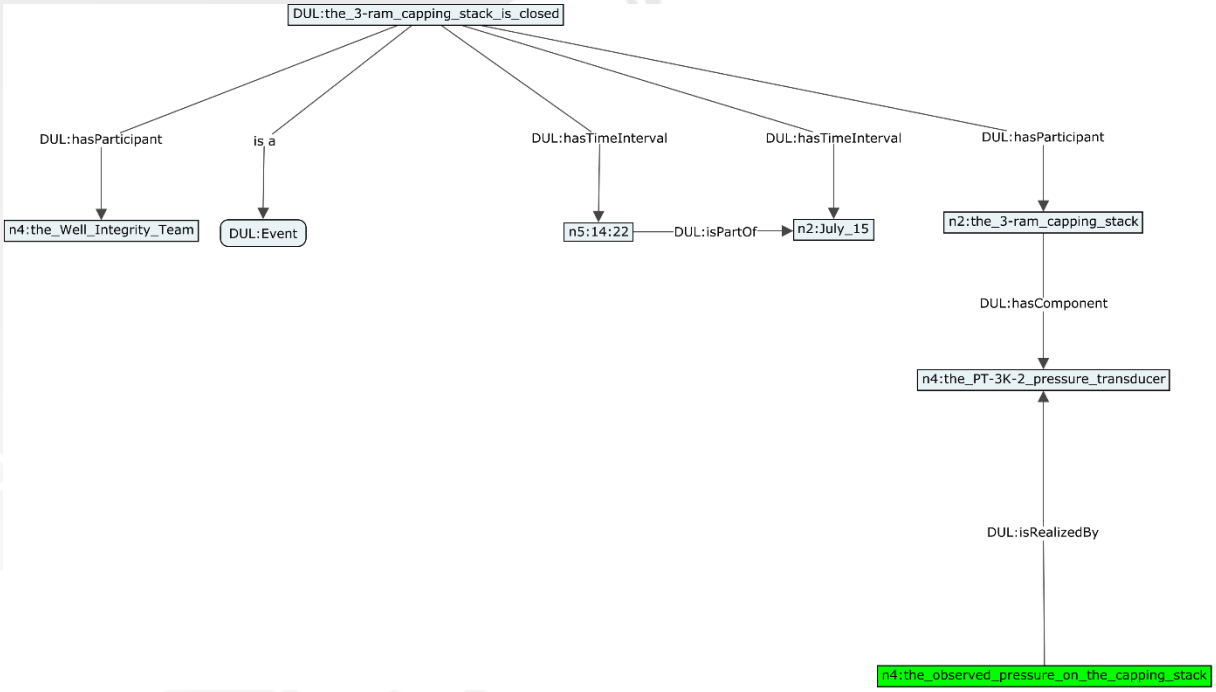


Figure 105: le *capping stack* est fermé.

La fermeture du *capping stack* a un effet direct sur le puits et sur le réservoir géologique : la pression sur les parois du puits, *pore pressure,* va augmenter car le puits est toujours actif et l'effluent très gazeux. Cette montée en pression, *pressure build-up*, est un indicateur utilisé comme un indice de l'état de santé du puits et des roches avoisinantes (Hickman *et al.*, 2012) et permet de déterminer l'intégrité du puits, c'est-à-dire d'évaluer sa capacité à supporter la pression et de détecter d'éventuelles fuites dans les formations géologiques environnantes. En effet, le puits ne pourra être gardé fermé si l'on n'est pas certain de son intégrité et notamment des disques de rupture[228] qui le composent. Si jamais cela ne devait être pas le cas, l'effluent pourrait s'infiltrer dans les roches avoisinantes, entraînant une éruption souterraine et l'huile pourrait venir suinter en surface (dans le fond de l'eau), « *une situation bien pire que le seul point de fuite par le* BOP *endommagé* » (McNutt, Chu, *et al.*, 2012).

Un modèle « *a priori* »[229] va donc être créé pour déterminer l'état d'intégrité du puits en fonction de la pression qui sera observée une fois le *capping stack* fermé. Le

226 Une fuite hydraulique sur une vanne du *capping stack* entraînera finalement un retard de 24 heures dans l'exécution de sa fermeture, effective le 15 juillet 2010.

227 On lit aussi à 14h25 dans le rapport (National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, 2011a, p. 165).

228 Équipements de sécurité installés à intervalles réguliers en profondeur et dont le but est justement de se rompre à une pression donnée pour protéger le puits d'une surpression interne.

229 Nous le nommons comme tel par commodité, il n'a pas de nom particulier.

réservoir géologique ayant une pression initiale estimée à 9000 psi (≈ 620 bar), la pression de référence, servant à considérer l'état d'intégrité du puits a été calculée à 7500 psi soit environ 517 bar, cette diminution étant une estimation de l'épuisement du réservoir (*depletion*) compte tenu des trois mois de fuite ininterrompue. A partir de cette valeur de référence, l'équipe d'intervention va envisager trois scenarii[230] :

1. si la pression est supérieure à 7 500 psi, le puits pourra être fermé pendant 48 h avec un faible risque, avec en prolongation, si la pression est supérieure à 8 000 psi, alors le puits n'est probablement pas endommagé et il est probable qu'il pourra être gardé fermé en toute sécurité pendant une plus longue période ;
2. si la pression est inférieure à 6 000 psi, alors la valeur indiquera que le puits a été considérablement endommagé et qu'il faudrait le rouvrir rapidement. Dans ce cas, on peut penser que le puits perd clairement de la pression quelque part sous le fond marin, probablement par l'éclatement des disques de rupture, et que des hydrocarbures s'infiltrent vraisemblablement dans les formations environnantes ;
3. enfin, entre 6 000 et 7 500 psi, le résultat sera ambigu, mais le puits pourrait être gardé fermé pendant 24 h, en supposant cependant qu'un écoulement au travers des disques de rupture de 20 000 barils maximum pourra être toléré sans dommage irréversible. Selon Hickman et al. (2012, p. 20270), si la pression se situe entre ces deux valeurs, les scientifiques et les ingénieurs seront confrontés à un dilemme, avec au moins deux explications possibles pour les résultats. L'une des explications est que certains des disques de rupture se sont rompus et que l'effluent s'infiltre lentement dans les formations géologiques environnantes. L'autre explication est que le réservoir fût plus épuisé que prévu (que le modèle le prévoit), entraînant une pression de fermeture moins élevée que prévu.

Le graphique[231] en page suivante résume la situation telle que l'équipe d'intervention la conçoit la veille de la fermeture du *capping stack* :

230 Les papiers de Hickman et al. (2012) et de McNutt et al. (2012) sont les sources scientifiques les plus fournies à ce sujet.
231 Source : (National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, 2011a, p. 165).

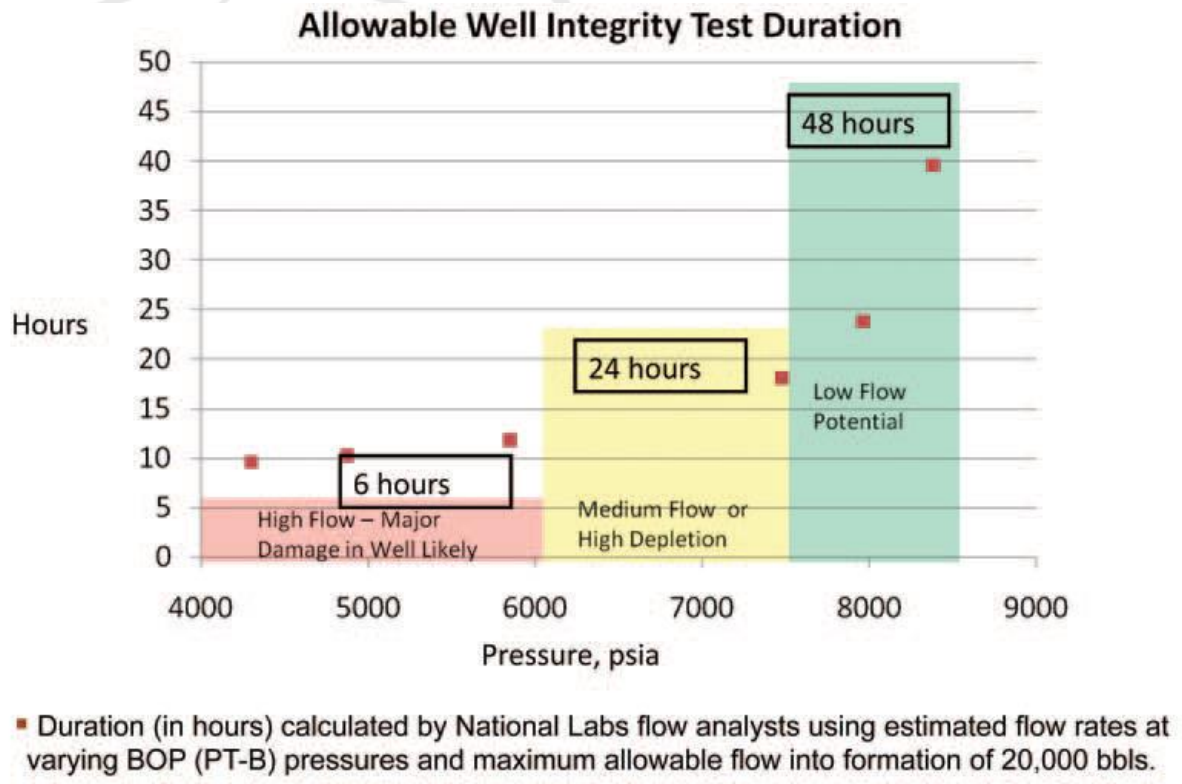


Figure 106 : hypothèses sur l'intégrité du puits en fonction des scenarii envisagés

Les deux premiers scenarii (zones rouge et verte) sont ceux qui permettent d'obtenir une bonne idée de l'intégrité du puits : la première possibilité est que le puits est intègre voire qu'il n'a pas été endommagé par l'éruption, il peut donc supporter la pression et rester fermé et la deuxième possibilité est plutôt le contraire, le puits a été terriblement endommagé et il faudrait le rouvrir au plus vite. Le degré de certitude est élevé. Le troisième scénario (zone jaune) est le plus redouté : l'état du puits ne pourra être déterminé avec certitude et il faudra le rouvrir au bout de 24 h si aucun changement n'a été apporté. Évidemment, ce dernier scénario n'est pas souhaité car il laisserait dans cette situation l'équipe d'intervention dans l'incertitude et compliquerait d'autant plus la suite des évènements.

Ce modèle est donc développé avant la fermeture du puits pour déterminer la marche à suivre pour la *WIT* quant au maintien fermé ou non du puits et la possibilité ultérieure de le « tuer »[232] définitivement.

C'est avant tout un problème de certitude auquel la *WIT* est confrontée : les scénarii rouge et vert, chacun à leur manière, ont des conséquences bien plus simples à anticiper que le scénario jaune, qui laisse la place au doute et ne permet pas de cerner la situation. Le graphe ontologique suivant montre le parallèle qui est construit dans le modèle entre l'intégrité du puits et la zone de pression de fermeture de ce dernier. De

232 Tuer un puits est une expression utilisée pour décrire l'opération qui permet de reprendre le contrôle d'un puits rentré en éruption d'une part par l'injection de grands volumes de boue pour rééquilibrer la pression dans le réservoir géologique puis d'autre part par l'injection de ciment pour sceller définitivement le puits.

là, découlent les décisions à prendre (non illustrées ici) une fois la pression observée stabilisée dans une zone ou dans une autre.

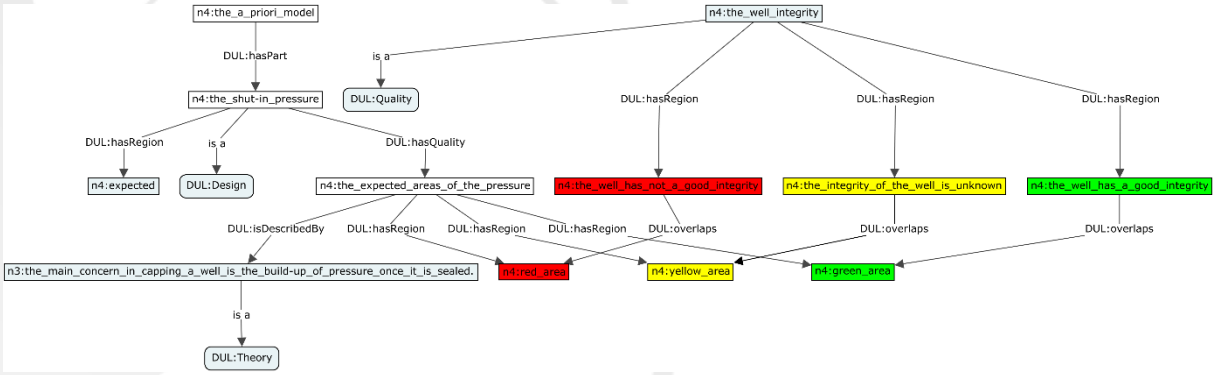


Figure 107 : le modèle *a priori*

Une campagne de mesure est menée pendant la fermeture et une fois le *capping stack* fermé. Le graphique suivant[233] (*pressure plot*) illustre la montée en pression, par étapes successives, du réservoir (et du puits) lors de la fermeture du *capping stack* :

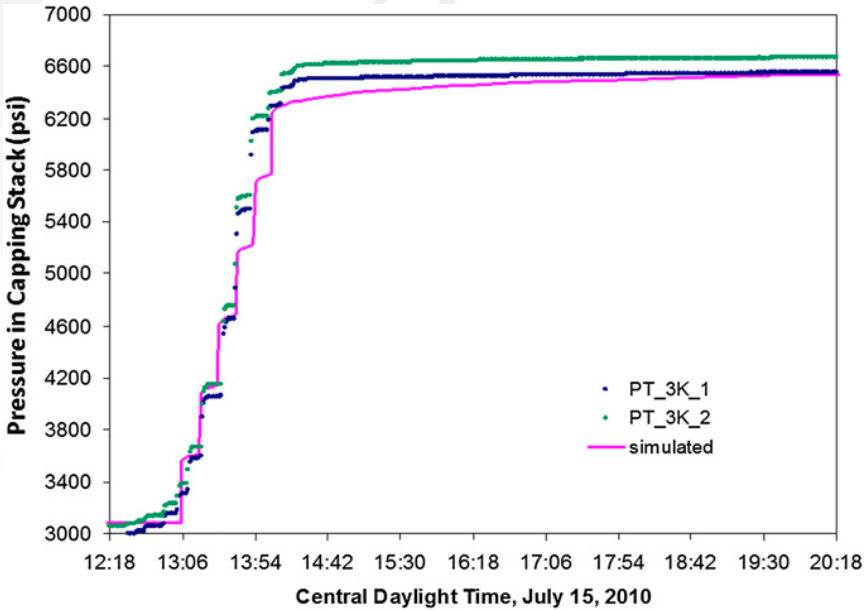


Figure 108 : la montée en pression du réservoir au fur et à mesure de la fermeture du puits (Hickman et al., 2012, p. 20270)

Les mesures sont effectuées par deux capteurs de pression (pour la redondance et dont la précision est de l'ordre de 2 psi), PT-3 K-1 et PT-3 K-2, qui ont été installés sur le *capping stack* à la demande des scientifiques de la *GLST*. En effet, avant l'accident, seul un manomètre à aiguille dont la précision est de l'ordre de 400 psi était installé sur le BOP et rendait difficile toute lecture et interprétation. Nous allons grandement revenir à la trace violette (la simulation) dans la suite de ce chapitre. Faisons abstraction pour le moment.

Il apparaît assez vite aux yeux de l'équipe d'intervention que la pression n'atteindra jamais les 7500 psi « rassurants » à propos de l'intégrité du puits (Hickman

233 Extrait de (Hickman *et al.*, 2012, p. 20270).

et al., 2012, p. 20270). La pression se stabilisera autour de 6600 psi. Cette valeur, au milieu du « nulle part », fait utiliser le terme *purgatory*[234] à un membre de l'administration[235]. Ce « lieu topologique de pénitence » peut être représenter de la sorte :

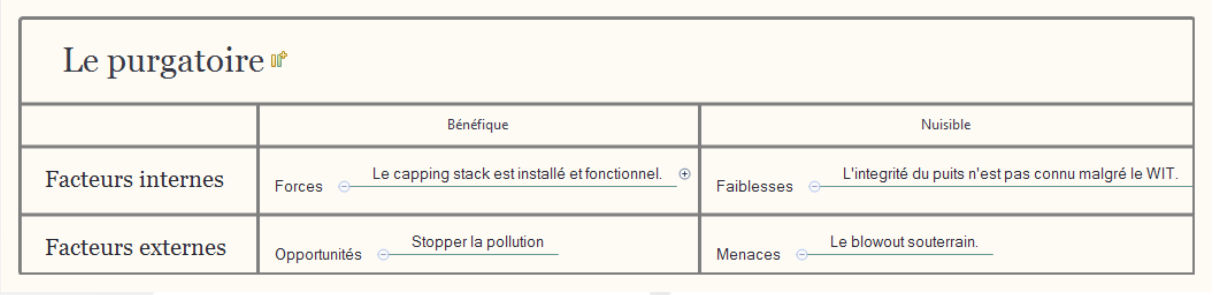


Figure 109 : le purgatoire

Des dissensions vont émerger au sein de l'équipe d'intervention. Il semble, à l'exception du personnel de BP, que personne ne souhaite garder le puits fermé. Richard Garwin, le père de la bombe à hydrogène US, disciple d'Enrico Fermi, alors membre de la *GLST*, était déjà un farouche opposant à la fermeture du puits et se demandait s'il n'était déjà pas trop tard.

Il faudra que l'amiral Cook, alors *Federal On-Scene Coordinator*[236] rappelle que cette situation, bien que non désirée, a été envisagée dans la procédure du test d'intégrité et que l'ensemble des intervenants était alors d'accord pour garder le puits fermé 24 heures maximum le cas échéant. L'équipe d'intervention suivra donc la procédure élaborée avant la fermeture du puits. Faute de changement, le puits sera rouvert dans 24 heures. Nous illustrons à la suite l'ontologie du « purgatoire ».

234 « Selon la Tradition, lieu (étant représenté souvent comme enflammé) où les baptisés, morts en état de grâce , mais non entièrement purifiés par la pénitence des traces de leurs péchés, achèvent leur purification avant la vision béatifique » Définition A donnée par le CNRTL (*PURGATOIRE : Définition de PURGATOIRE*, no date).

235 Voir (National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, 2011a, p. 165).

236 Chef des opérations sur site, pouvoir délégué du *NIC*.

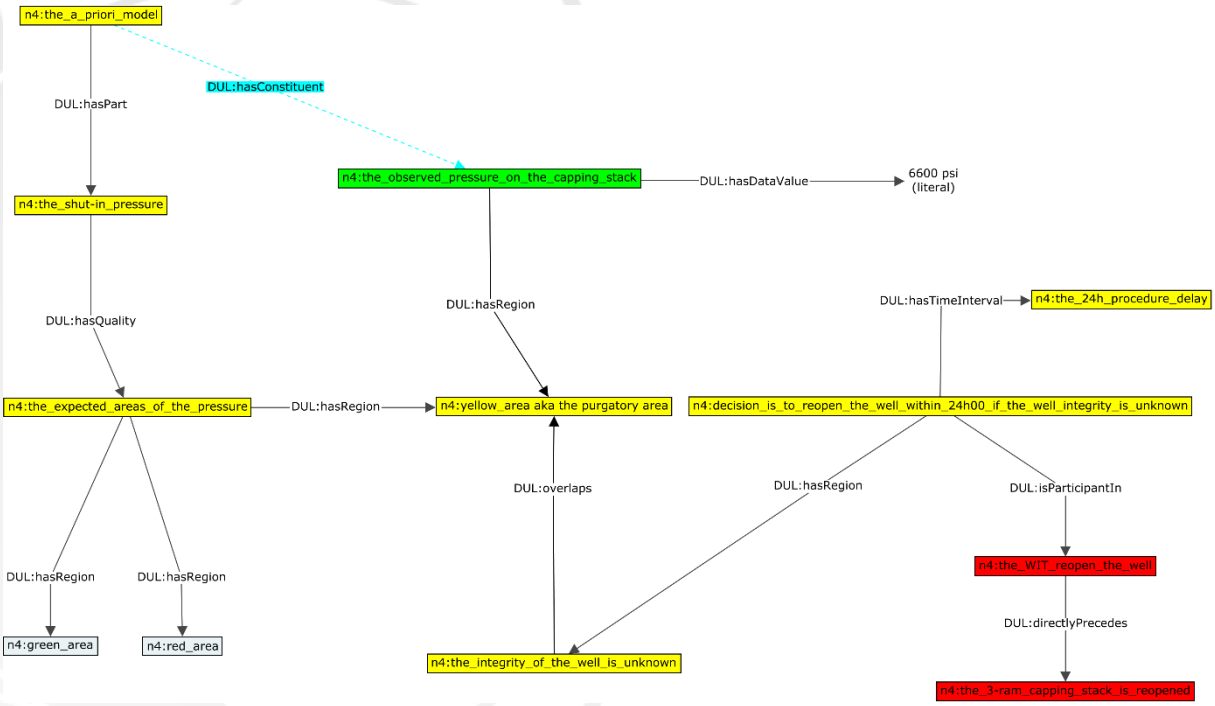


Figure 110: l'ontologie du « purgatoire », le puits sera rouvert dans 24 heures.

Pour sauver le Golfe, il faut donc trouver le meilleur changement possible pour garder le puits fermé. L'objet central de ce graphe ontologique est l'évaluation de l'intégrité du puits, *via* la mesure de la pression de fermeture confrontée au modèle géologique « *a priori* ». Pour lire ce graphe, il faut comprendre que le « cheminement de pensée » de la *WIT*, une fois qu'elle constate la valeur de la pression au niveau du *capping stack*, est illustré en jaune. La pression à la fermeture « atterrit » dans la zone jaune, signifiant que l'intégrité du puits n'est pas certaine, et donc que la *WIT* devra rouvrir le *capping stack* dans les 24 heures s'il n'y a pas de changement et que la pollution continuera (les deux instances en rouge).

Or, nous le savons, cela n'a pas été le cas. Le *capping stack* a bien été maintenu fermé. L'explication qui fait consensus dans l'ensemble des documents que nous avons pu consulter est illustrée par le graphe ontologique en page suivante :

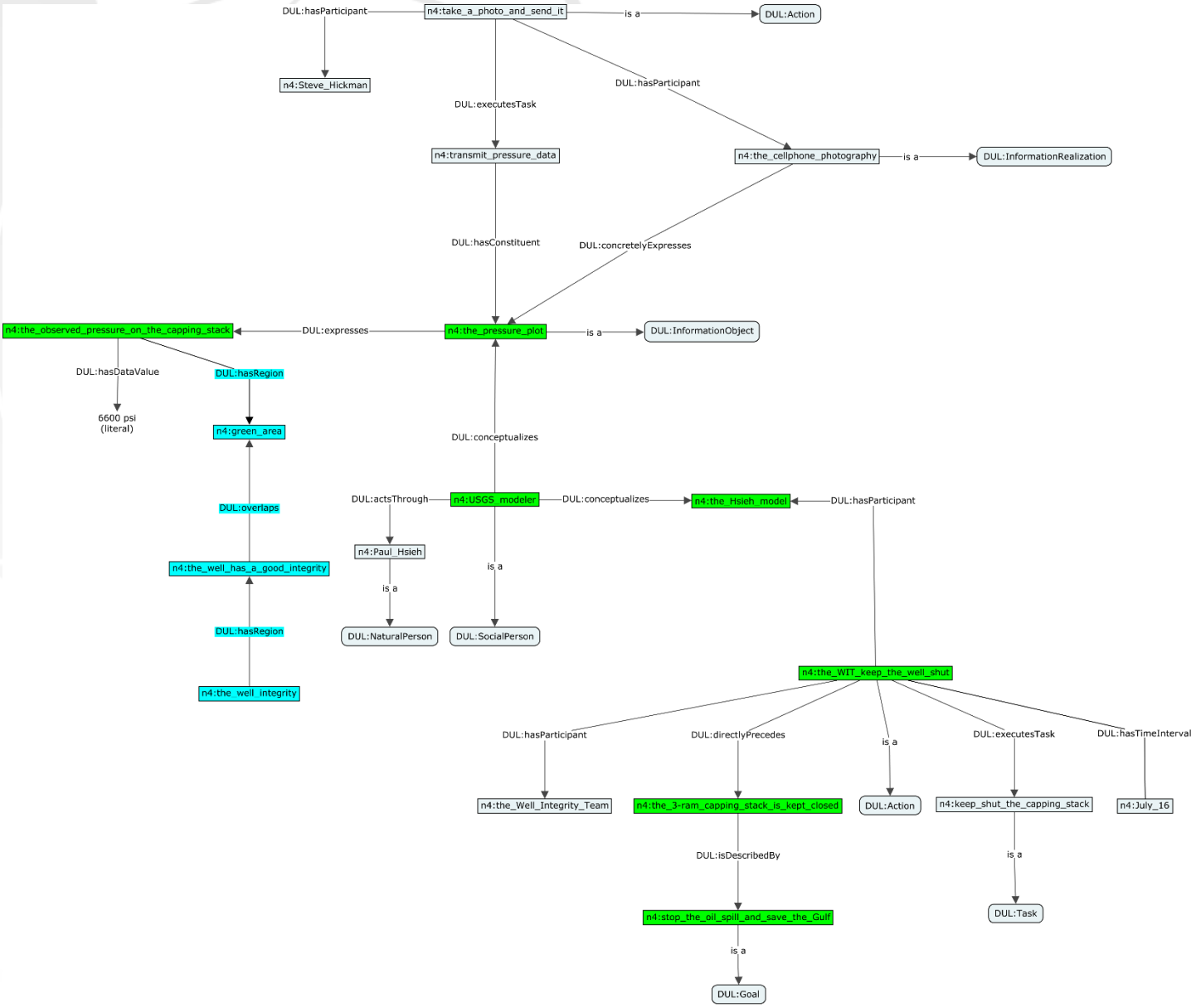


Figure 111 : le *capping stack* restera fermé le 16 juillet 2010, marquant la fin de la pollution.

L'explication donnée suit le cheminement en vert : la valeur observée de la pression va être soumise, par une habile transmission du graphe de l'évolution de cette dernière (« *take a photo and send it* »), à l'expertise d'un géologue de l'USGS, le Dr. Hsieh. Celui-ci, alors qu'il n'est pas présent au centre de gestion de crise de Houston ce soir-là, va travailler pendant la nuit du 15 juillet à partir de ce relevé de pression. Il proposera le lendemain matin un modèle susceptible d'expliquer la valeur observée comme étant compatible avec celle d'un réservoir d'une certaine forme (dont il fait l'hypothèse sur la base du test de plusieurs formes) et donc *in fine* d'un puits en bonne intégrité. Cette explication fera consensus et la *WIT* décidera donc de maintenir le puits fermé sans que cela ne provoque de *blowout* souterrain et mettra ainsi fin à la pollution. Le modèle de Hsieh « a créé la branche bleue » dans l'ontologie.

Foncièrement, dans le récit, nous passons en une nuit à une interprétation contraire de ce qui avait été pensé la veille ; l'approche est binaire : le 15 juillet, la *WIT* doit rouvrir le puits, le 16 juillet, elle le laissera fermer. Ce changement est d'autant

plus surprenant si l'on met en perspective l'enjeu de situation du Golfe du Mexique avec la manière dont le renversement total du consensus scientifique du départ s'est effectué. Dans les deux graphes ontologiques que nous venons de présenter, cela se traduit par l'observation (au sens de l'ontologie) de la même instance « *the observed pressure on the capping stack* », mais à des périodes temporelles différentes (que l'on peut donc comprendre comme des points de vue selon *DOLCE*). Ce n'est pas la temporalité en tant que telle qui nous intéresse, mais bien deux interprétations diamétralement opposées du même phénomène (représentée par une pression) qui sont proposées en 24 heures. En formalisant la connaissance par l'utilisation d'une ontologie, nous mettons en évidence, pour la communauté scientifique, la nécessité de se pencher sur cette instance-clé qui représente la pression observée au niveau du *capping stack*. Notons toutefois que ce type d'analyse serait grandement facilité par des modes de représentations de type chronologique. Une idée de développement pourrait être une représentation graphique pouvant permettre de visualiser la construction ontologique de la connaissance en rapport avec une structure chronologique (amenée elle aussi par l'ontologie). En « fixant » une chronologie, à la manière d'un axe de référence statique, on ferait « tourner autour » la connaissance et nous pouvons penser que cela aiderait à raisonner, notamment en matière de causalité et faciliterait les découvertes.

Il s'agit donc de comprendre comment nous sommes passés d'une interprétation à une autre. En effet, si l'on s'arrête à ce niveau de connaissance, on ne peut pas expliquer pourquoi il a été rendu possible de garder le puits fermé, ni d'évaluer le mécanisme à la base du changement d'un modèle pour un autre (du modèle « *a priori* » à la simulation de Hsieh). Et pourtant, les rapports d'enquêtes qui mentionnent cet épisode ne vont pas au-delà de la simple description de ce changement (brièvement d'ailleurs), et de la même manière, les papiers de science qui évoquent cette période ne donnent guère plus de détail à ce sujet. Il va falloir aller chercher plus en profondeur pour expliquer cette bascule ontologique de la valeur de cette pression.

4.2.2 A la recherche de l'explication manquante

Nous considérons la valeur de la pression observée comme la clé de la compréhension de ce retournement de situation. Il faut d'abord comprendre comment les modèles géologiques sont effectués car ils sont présentés comme les fondations pour les interprétations par la *WIT* des valeurs observées. Commençons par le modèle « *a priori* ». On pourrait écrire la chose suivante :

1. la phase *A*, phase d'élaboration du modèle[237] qui repose sur des hypothèses *a priori* sur le réservoir géologique avant que le *capping stack* ne soit effectivement fermé ;

puis,

237 Il s'agit en fait de combinaisons de modèles. Cependant tous ces modèles ont été élaborés par BP et ont été revus et commentés par la *GLST*. Voir l'entretien avec Paul Hsieh à ce sujet. Par la suite nous désignerons ces modèles par le singulier.

2. la phase *B*, la phase d'observation des pressions observées[238] au niveau du *capping stack* ;

et enfin,

3. la phase *C*, la phase d'interprétation des observations obtenues en *B* à la lumière de la modélisation conçue en *A*.

Le raisonnement est déductif. A la fermeture du puits, lorsque la pression s'établit dans la zone jaune, les 24 heures de la procédure sont considérées par l'administration comme une ultime chance de laisser une opportunité d'arriver à garder le puits fermé (Hickman *et al.*, 2012, p. 20270).

Sauf erreur de mesure, il n'est pas possible de changer les valeurs de la pression mesurée lors de la phase *B*. *B* est « factuel ». La situation, bien qu'envisagée, n'est pas du tout désirée. L'échappatoire proposée face à cette incertitude est de rouvrir le puits. Bien que cette décision soit la plus rationnelle[239] à cet instant, elle ne semble convenir à personne. La *WIT* envisage la possibilité que la pression observée puisse finalement être de bon augure pour garder le puits fermé. Pour cela, les 6600 psi doivent correspondre à un puits intègre : c'est la nouvelle interprétation (*C'*) que la *WIT* souhaite donner aux observations de la pression, et c'est ce que va justifier l'analyse de Hsieh.

Hsieh, que nous avons donc rencontré, est un géologue hydrologue dont les compétences sont très reconnues par ses pairs et par l'administration[240]. Et, particulièrement pendant cette nuit décisive du 15 au 16 juillet puisqu'il sera sollicité alors qu'il est rentré chez lui, à Menlo Park, pour s'occuper de son père âgé. Hsieh va proposer un modèle susceptible de montrer qu'il est possible que la pression observée soit compatible avec un puits en bonne intégrité. A la lecture du rapport de Hsieh (2010), bien au-delà des subtilités géologiques qui nous intéressent moins, c'est la méthode de modélisation qui nous a frappés. Nous avons appris qu'il est pleinement possible de « remonter » le fil de l'observation d'une valeur de pression de réservoir géologique pour déterminer la forme et les caractéristiques de ce réservoir, rendant également possible de pouvoir estimer l'intégrité d'un puits foré dans ce réservoir. Cette méthode d'analyse est appelée « *inverse modeling* » et est décrite comme : « *le processus par lequel les entrées d'un modèle informatique d'un système naturel ou artificiel sont ajustées jusqu'à ce que les sorties du modèle, telles que la hauteur de chute hydraulique, le flux d'eau souterraine ou les concentrations de solutés,*

238 Mesures redondantes prises par les capteurs PT_3K_1 et PT_3K_2. Nous simplifierons de la même manière en utilisant le terme au singulier.

239 Telle que la rationalité est décrite dans les papiers et de l'idée qu'en ont les scientifiques de la *GLST*. Très grossièrement la rationalité consiste à « mettre des chiffres » (« *rationales* ») sur le monde pour le comprendre et créer un canal de communication au sein de l'équipe d'intervention pour finalement appuyer une approche très conservatrice compte tenu des enjeux. Nous employons ce terme comme un concept « local » et il n'est pas question ici de faire débat. Il faut lire notre interview de Steven Chu à ce sujet et l'opposition qu'il fait avec la « *cowboy mentality* » qu'il attribue à certains ingénieurs de BP.

240 Source : (*USGS Scientist Honored with Prestigious Federal Employee of the Year Medal for Role in Ending Deepwater Horizon Oil Spill*, no date).

correspondent aux observations de terrain. »(Dahlstrom and Carter, 2013, p. 162). L'utilisation de logiciels de modélisation géologique spécialisés tels que PEST[241] permettent d'effectuer cette analyse. Dans le cas du travail de Hsieh, on pourra parler de « *model calibration* » (Introduction 1-2 'PEST Model-Independent Parameter Estimation User Manual, 5th Edition.', no date), Hsieh parlant d'« *History Matching* » dans son rapport. L'objet est *in fine* de déterminer les caractéristiques les plus « plausibles » du réservoir et de voir si elles peuvent amener à penser que le puits est intègre. Il utilise la pression mesurée par le capteur PT-3 K-2 comme valeur d'entrée pour la modélisation inverse.

Le graphe ontologique en page suivante représente la « confrontation des modèles » entre le modèle *a priori*, basé sur des hypothèses et selon un raisonnement déductif, et le modèle inverse, qui construit une théorie explicative (de) à partir de l'observation :

241 Source : ('PEST Model-Independent Parameter Estimation User Manual, 5th Edition.', no date).

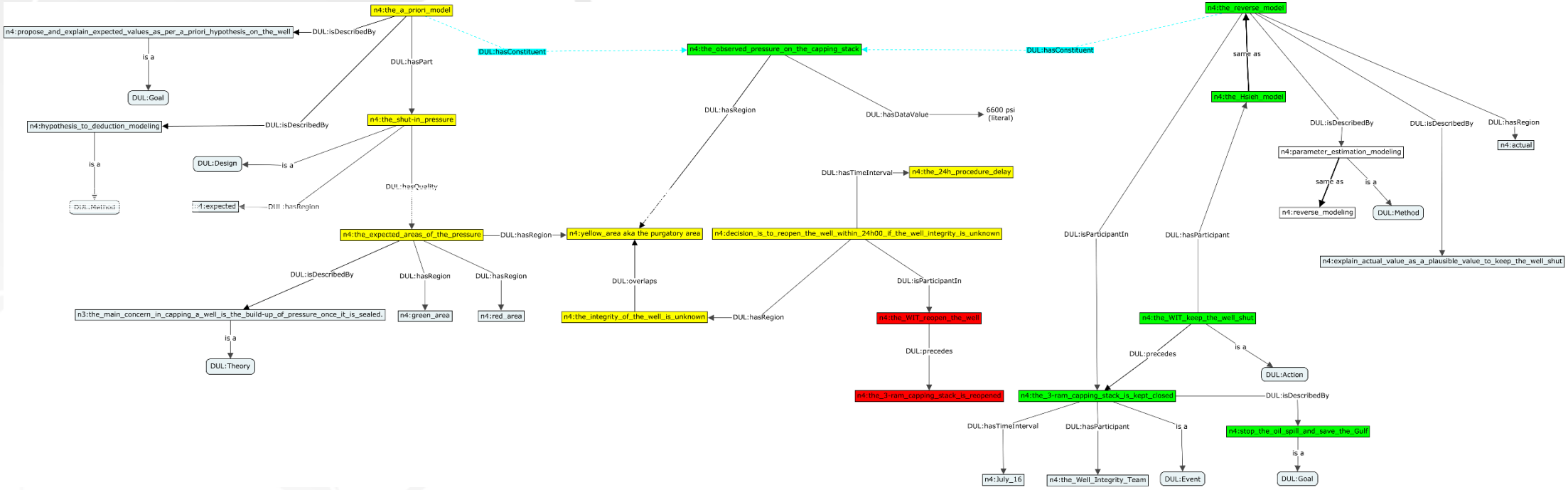


Figure 112 : la confrontation des modèles

Avec ce graphe ontologique, nous mettons en évidence que le modèle hypothético-déductif utilisé en premier bloque la situation car il ne peut pas proposer de solution autre que celles contenues déjà dans ces hypothèses (cheminement jaune puis rouge), tandis que le modèle inverse permet de trouver une solution théorique qui satisfasse la plausibilité de la réalité de cette solution et permet de débloquer la situation (cheminement vert). Néanmoins, cela reste au modélisateur d'évaluer la plausibilité des solutions théoriques proposées. Hsieh va donc « jouer » sur différents paramètres qui caractérise un réservoir géologique[242] et proposer un modèle à la *WIT* le matin du 16 juillet. Le modèle produit une simulation de la pression (qui correspond à la sortie du modèle inverse), à comparer à la pression observée. Nous reprenons le graphique (*pressure plot*) précédemment montré :

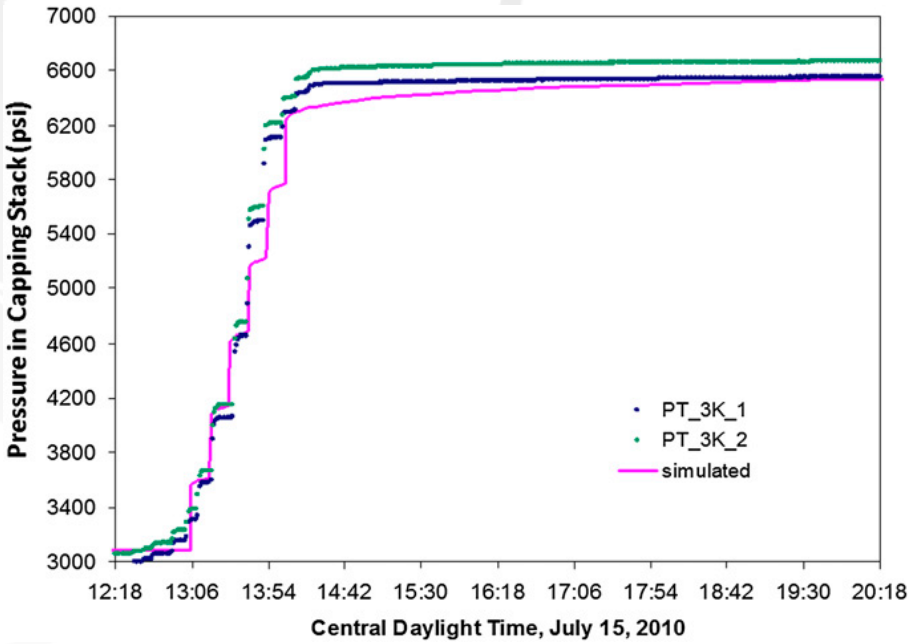


Figure 113 : la pression simulée par le modèle ajusté de Hsieh

> « *L'étroite correspondance entre les pressions observées et les pressions simulées a montré qu'il existait un scénario raisonnable dans lequel le puits Macondo avait une intégrité totale (c.-à-d. sans aucune fuite après sa fermeture), mais que le réservoir avait été considérablement épuisé durant l'éruption.* » (Notre traduction Hickman et al., 2012, p. 20270)

La *WIT* décidera, après une ultime réunion de concertation entre les ingénieurs de BP, la *GLST* et Paul Hsieh depuis Palo Alto, de garder le puits fermé. Les jours suivants vont confirmer le bien-fondé de cette décision. Force est de constater qu'il n'y a pas eu de conséquence désastreuse (comme un *blowout* souterrain).

4.2.3 Causalité contrefactuelle et expression dans *DOLCE*

L'analyse de Hsieh a donc pour but de « *[…] voir si les pressions mesurées au niveau du capping stack immédiatement après sa fermeture pourraient être expliquées sans invoquer de fuites souterraines* » (Hickman *et al.*, 2012, p. 20270). Les scientifiques sont à la recherche d'un « *[…] argument plausible pour permettre à*

242 Voir les tableaux 1 et 2 en annexe de son rapport pour les paramètres en question.

la WIT de continuer au-delà de la période initiale de 24 heures » (McNutt, Chu, *et al.*, 2012, p. 20224). Il faut donc créer un univers théorique où la pression observée correspondrait à un réservoir géologique où le puits serait intègre et permettrait sa fermeture en sécurité. C'est un raisonnement surprenant où l'on part de l'observation phénoménologique (B) pour tenter de trouver une théorie nouvelle (A') permettant une nouvelle interprétation (C') de cette observation, plus conforme à ce qu'on souhaiterait qu'elle soit. L'enjeu est donc de trouver un monde plausible dans lequel (C') devient possible, et pourra, sur la base des mesures « objectives » de pression, être considérée comme « vraie ».

C'est la découverte d'un tel monde possible qui « explique » la bascule entre les deux situations que nous avons présentées auparavant. Cette explication n'apparaît pas en tant que telle dans les documents que nous avons pu consulter et la formalisation ontologique nous a permis d'y voir plus clair et d'orienter nos recherches par une confrontation « brutale » entre deux situations radicalement différentes à propos du même objet observé.

Si nous souhaitons formaliser le raisonnement qui a été effectué par l'équipe d'intervention, nous pouvons dire maintenant qu'il existe au moins un modèle de réservoir avec un puits intègre dont les caractéristiques théoriques amènent à observer, dans les conditions du moment, la pression en tête du puits MC 252-1 telle qu'elle fût mesurée. D'ailleurs, ce « premier modèle de Hsieh » sera ajusté par la suite au fur et à mesure des observations (dans les jours qui suivirent la fermeture) pour passer d'un réservoir grossièrement cubique à un réservoir plus parallélépipédique (le « deuxième modèle de Hsieh »).

Nous faisons remarquer que Hsieh aurait pu ne jamais trouver de modèle de réservoir qui corresponde aux mesures parce que, dans tous les univers possibles, il y aurait pu n'y avoir aucune construction théorique qui satisfasse à la fois la réalité de cette pression et la capacité (nécessité ou volonté) à vouloir interpréter cette dernière comme « un reflet » d'un réservoir avec un puits intègre. Et la *WIT* aurait dû rouvrir le puits.

Hsieh a trouvé un point de convergence remarquable entre une réalité plausible et une croyance « imposée » qu'il a rendu réel. C'est un raisonnement contrefactuel qui a permis de maintenir le puits fermé et de stopper la pollution. Il est intéressant de noter que McNutt et al. utilisent bien le terme de « *plausible* » et nous ne pouvons nous empêcher de faire le rapprochement avec Griffin, citant Hawthorn, à propos du monde contrefactuel et de la plausibilité de ce dernier :

« *[…] Hawthorn impose des conditions […] à la plausibilité du raisonnement contrefactuel : les mondes possibles devraient (a) se baser sur le monde réel comme il était autrement connu avant de faire une hypothèse contrefactuelle* [un monde possible]*, (b) ne pas exiger de* « *débobiner le passé* » *et (c) ne pas trop* « *bouleverser* » *ce que nous comprenons par ailleurs des acteurs et de leurs contextes* » (Notre traduction 1993, p. 1102). Le raisonnement suivi a rempli ces trois conditions : (a) le monde réel est bien représenté par les 6 600 psi de pression en tête de *capping stack*, cette valeur étant connue avant d'émettre l'hypothèse d'un puits intègre ; (b) il n'y a aucun passé à « débobiner » puisque c'est la première fois à laquelle est confrontée

l'équipe d'intervention, il n'y a donc aucune hypothèse qui aurait pu être formulée à ce propos ; (c) les acteurs (la *WIT* et le Dr. Hsieh) ainsi que le contexte ne sont pas du tout bouleversés par le « monde possible » espéré par la *WIT* et recherché par Hsieh.

Or, nous avons vu auparavant qu'il est assez simple d'exprimer une causalité de type implication ou prérequis dans *DOLCE* par le biais de la famille de prédicats *hasPrecondition*. On l'a vu aussi, ces prédicats héritent des familles *precedes* ou *follows* qui imposent explicitement la notion de séquence. Cependant, avec la causalité contrefactuelle, il ne s'agit pas de raisonner en séquence, mais plutôt en élaborant des « mondes possibles » dont il faut évaluer leur plausibilité à se substituer à l'encontre des faits pour tenter d'expliquer la survenance d'un évènement. Il s'agit selon nous de créer « des mondes parallèles », imaginaires, et qui cohabitent avec le monde réel, jusqu'à qu'ils y soient confrontés au détour d'un questionnement contrefactuel : Hsieh répond à la question : est-il possible que le puits soit intègre quand même malgré la valeur de la pression observée ? Hsieh va rechercher un « monde parallèle » où cela serait le cas et qui serait plausible de substituer au monde réel.

Exprimer ce mode de raisonnement à l'échelle atomique (le *triple*) n'est pas possible à l'heure actuelle dans *DOLCE* comme c'est déjà le cas pour exprimer l'implication ou le prérequis. La représentation de l'explication que nous venons de donner nécessiterait une population conséquente de *DOLCE*, mais il n'est dit que cette représentation formalise correctement l'explication. Il est fort probable qu'il faille modifier l'ontologie (en termes de classes et de prédicats) pour tenter d'apporter le formalisme nécessaire à la contrefactualité. Ce travail dépasse le cadre strict de cette thèse, mais il est essentiel à signaler car nous touchons le bord de l'expressivité de l'ontologie face à un « concept » qui est difficile à appréhender par ce biais. De plus, notre algorithme de recherche de causalité dans les textes devrait être en mesure de détecter de tels liens de causalité, s'il est correctement entraîné sur une base de données qui identifie les liens de causalité contrefactuelle (ce qui n'est pas le cas aujourd'hui dans les bases de données d'entraînement, puisqu'au contraire les critères de contrefactualité servent à identifier un lien comme « causal » ; le raisonnement contrefactuel n'est donc pas identifié comme tel). Mais le lien entre cet algorithme et l'ontologie, pour être effectif, devra reposer sur une modalité d'enregistrement de la causalité contrefactuelle dans l'ontologie.

4.3 Discussion et limitation des résultats

Nous discutons maintenant de nos résultats et présentons les principales limites.

4.3.1 Les limites de notre étude de l'accident *Deepwater Horizon*

L'analyse de l'espace sémantique de l'unité lexicale « accident » a permis de ramener au centre de la discussion la définition de l'accident, qui semblait quelque peu figée, autant dans son acception commune que dans les *safety studies*. Cela devrait susciter le débat puisque nous avons montré qu'un cas comme *Deepwater Horizon* ne saurait être étudié convenablement sans aller au-delà des modèles d'accident existants et de la nécessité de convoquer le cadre conceptuel de l'ingénierie en situation extrême

développé par les chercheurs du CRC. Notre recherche s'est volontairement limitée à cet axe-là, mais nous avons bien conscience qu'un cas comme *Deepwater Horizon*, tel que nous souhaitons qu'il soit compris maintenant, mériterait de bien plus amples investigations par le prisme d'autres disciplines et *via* d'autres approches.

Concernant les données de cet accident, nous nous sommes exclusivement intéressés à l'écrit, aux données textuelles, c'est-à-dire le texte écrit en langage naturel (très majoritairement en anglais) et l'environnement quantitatif afférent le cas échéant. Nous ne nous sommes pas foncièrement penchés sur d'autres types de données. Nous pensons notamment aux données graphiques (photos, dessins, films…) et données audio (enregistrements sonores, audiences…). Cette limite s'est imposée presque naturellement parce que nous supposons les données textuelles plus importantes en volume et plus riches en connaissances à forer.

Ensuite, autant nous pouvons considérer aujourd'hui que nous avons une certaine expertise dans le secteur d'activité *oil & gas*, autant il serait prétentieux de penser que cette expertise n'eut pas de limite et notre vision du cas *Deepwater Horizon* est inévitablement bornée par ces limites. Il en va de la crédibilité de notre démarche experte. De la même manière, mais à l'opposé, cette thèse a été aussi un conséquent travail d'apprentissage de nombreuses disciplines, notamment la linguistique et l'algorithmie, peu enseignées (c'est un euphémisme) dans les écoles d'officiers de la marine marchande. Ces travaux ont été aussi un aller-retour permanent entre l'atteinte de paliers de maîtrise d'un concept, d'un outil, etc. et leur mise en œuvre pour aller à la découverte d'autres concepts…

Enfin, deux limites matérielles, mais extrêmement chronophages ont été la très forte nécessité de trouver le moyen de créer de l'interopérabilité entre les nombreux logiciels utilisés et la multitude des formats de données et le nettoyage[243] quasi-indispensable des données à traiter. Nous pensons qu'il s'agit d'un véritable problème dans cette discipline que sont les *data science*. Par nécessité, ne sachant pas coder, mais aussi pour ne pas réinventer la roue à chaque fois, nous avons choisi le parti-pris de systématiquement tenter de trouver un logiciel (ou un programme ou un « bout de code » exécutable) existant pour pouvoir traiter nos données et apporter les résultats que nous avons. Il va sans dire que nous recherchions en priorité du code *open-source* sourcé et entretenu. Cette démarche a été la source de belles découvertes de logiciels de qualité développés par cette communauté hybride que sont les *data scientists*, mais cela a été aussi le fruit de nombreuses frustrations et de perte de temps. Puis vient l'enjeu de pouvoir faire opérer ces logiciels entre eux par le biais d'un fichier de données avec les mêmes problèmes que nous venons de présenter. Le nettoyage des données a lui aussi coûté beaucoup de temps en plus d'exiger beaucoup de méticulosité.

4.3.2 Les limites des ontologies pour la formalisation des connaissances et le traitement de la causalité

La première limite à nos travaux est l'ontologie en tant que telle. Nous l'avons vu tout au long de ces travaux, il faut reconnaître que les ontologies sont des « objets »

243 Source : https://www.techopedia.com/definition/1174/data-cleansing

avec une courbe d'apprentissage ardue dans leur conception et que leur implémentation à des fins d'utilisation demande aussi un effort qui peut éventuellement freiner leur déploiement ou, plus prosaïquement, les « dévaloriser » vers des bases de données plus « accessibles » et ainsi perdre toute leur puissance conceptuelle et les opérations logiques qui en découlent. Guarino (1997, 1995 ; 2009) est un fervent partisan de l'indépendance des connaissances face aux problèmes qu'elles sont censées résoudre ou des domaines qu'elles sont censées représenter. L'ontologie doit rester libre du domaine de connaissance et indépendante du problème qu'elle aide à résoudre. Nous nous inscrivons dans cette école de pensée ; nous comprenons aisément qu'une ontologie puisse être très dépendante d'un contexte quelconque (domaine, application…), mais alors dans ce cas, elle ne saurait répondre aux critères d'universalité et perdrait probablement de sa capacité à faire émerger « les propriétés les plus générales de l'être »[244] et donc à pouvoir être « portée » dans d'autres domaines de connaissances et *in fine* perdre son pouvoir de représentation[245].

Poli (1999, p. 3) expose les deux oppositions qui existent dans la conceptualisation des ontologies à savoir :

- une opposition dans l'orientation entre objet et concept dans ce que l'ontologie doit « configurer ». L'auteur considère que le regard à porter sur les concepts est un problème épistémologique et non ontologique ; il s'agit de réfléchir d'abord à la manière de construire des connaissances et non pas de savoir ce qu'elles sont en tant que tel ;
- une opposition de dépendance (ou d'indépendance) de l'ontologie au « domaine » qu'elle est censée représenter.

Il faut comprendre que les deux controverses sont toujours présentes (19 ans après) et que certains considèrent une ontologie indépendante de son domaine comme une chimère[246]. Dans une note de bas de page, Teulier et al. assurent que les ontologies qui fonctionnent réellement sont des ontologies d'applications « *spécialisées et influencées par la tâche considérée* » (2005, pp. 14–15).

Un autre risque est celui d'idiosyncrasie. Typiquement, à propos de réutilisation des ressources disponibles (« ne pas réinventer la roue »), Smith (2006, p. 4) reproche aux concepteurs de la taxinomie de l'ISO 15926 de ne pas avoir fondé leur travail sur la théorie des ensembles classique. Et, en tentant une appropriation dans leur manière d'appréhender les ensembles, Smith fait remarquer que :

> « *Mais parce que cette ressource [la théorie des ensembles entre autres] a apparemment été ignorée par les développeurs de l'ISO 15926, le résultat est un charabia, que personne (ou du moins : personne en dehors de la communauté de modélisation des données de l'industrie pétrolière et gazière) ne ressentirait normalement le besoin d'utiliser, et des définitions qui ne sont, en aucun cas, accessibles au grand public* » (Notre traduction 2006, p. 4).

244 Voir la définition de l'ontologie donnée par le CNRTL.
245 Même si, bien entendu, cette notion « d'universalité » doit, selon nous, être interprétée à la lumière d'un imaginaire social donné, éventuellement largement partagé.
246 La chimère évoquée dans le chapitre 2.

Nous devons avouer que, même en ayant la prétention de penser appartenir à la communauté « *oil and gas* » et commençant à maîtriser un tant soit peu la représentation des connaissances, nous n'avons pas compris non plus où voulait en venir l'ISO 15926 à propos de sa taxinomie.

Nous venons de présenter les limites aux ontologies. Il s'agit finalement plus de limites liées aux concepteurs qu'aux ontologies elles-mêmes. Nous proposons une illustration synthétique des limites :

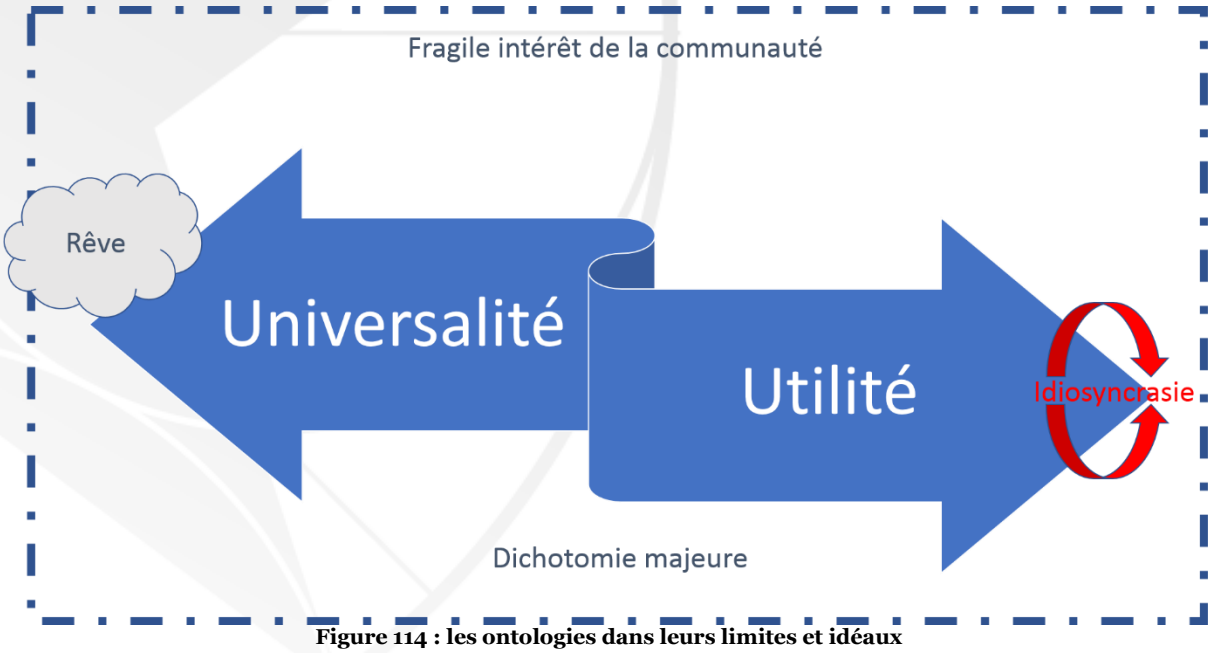


Figure 114 : les ontologies dans leurs limites et idéaux

La difficulté majeure de « l'ontologue » n'est pas de rester à l'équilibre entre les deux tensions, mais bien d'éviter de sombrer d'un bord ou de l'autre, poussé par l'idéal d'universalité ou d'utilité. Il faut revenir aux quatre piliers exposés dans le chapitre 2 dès lors que l'on souhaite travailler avec les ontologies à des fins de représentations de connaissances.

Plus proche de nos travaux, une autre limite afférente à l'utilisation des ontologies est le manque d'outils de visualisation dignes de ce nom. Nous ne pouvons pas dire pour le moment qu'il soit aisé et agréable de « naviguer » dans une ontologie écrite en *OWL*. Nous sommes encore très loin de ce qui peut être fait avec les taxinomies biologiques et des outils comme *LifeMap*[247]. Nous avons bien heureusement trouvé une solution palliative avec l'utilisation d'une version ajustée de *CMAP Tools*, mais qui ne prend malheureusement pas en compte les annotations *OWL* dans le tracé des graphes. Mais la plus grande limite à l'utilisation des ontologies reste le processus de population. Notre revue de littérature ne nous a pas permis de trouver un logiciel adapté et nous avons exposé toute la difficulté et les risques encourus à le faire à la main. Notre proposition d'algorithme *NER* présentée dans le chapitre 3 est une réponse ciblée. Cet algorithme n'adressera pas toutes les difficultés liées à la population d'une ontologie, mais tend à apporter un automatisme bienvenu, à la fois

247 Source : http://lifemap.univ-lyon1.fr/explore.html

parce qu'il fait économiser du temps, mais surtout parce qu'il appliquera systématiquement le même « travail de conceptualisation » au texte étudié, affranchi des biais cognitifs humains.

Un apport majeur de notre thèse est la proposition d'une machine qui analyse la causalité exprimée dans un texte et est capable de fournir un cheminement causal plausible comme réponse à la question pourquoi ? Nous avons montré comme il est plutôt aisé de décrire la causalité dans *DOLCE* tant que celle-ci est exprimée comme un prérequis ou une implication. Tant que cette causalité suit un schéma séquentiel dépendant d'une temporalité. Mais nous avons vu qu'il est beaucoup plus difficile si ce n'est impossible de représenter un raisonnement contrefactuel dans *DOLCE, a fortiori* quand ce dernier est déduit de la compréhension humaine d'un texte et non explicitement décrit dans un document. On touche ici une limite quant à l'adéquation entre l'engagement ontologique et le niveau d'interprétation d'un texte. Comme on l'a dit pour le processus de population, l'enjeu est de réduire systématiquement la connaissance à son niveau atomique, échelle élémentaire d'expressivité, qui permet le traitement humain et automatique. Or, un cas de raisonnement contrefactuel comme nous l'avons découvert et exposé dans la partie 2 de ce chapitre est un véritable défi ontologique à réduire de la sorte. De plus, c'est le lecteur humain qui met en évidence ce mode de raisonnement à partir de l'analyse du récit. Notre machine jouera le rôle d'un éclaireur (*pathfinder*) en proposant des cheminements causaux grâce à la causalité exprimée dans les documents, mais encore faut-il que cette dernière le soit ! Et même avec toute la connaissance disponible et explicitement écrite dans un texte, la question de la conceptualisation du raisonnement contrefactuel dans l'ontologie reste ouverte. Nous ne sommes pas en mesure d'apporter une réponse aujourd'hui dans *DOLCE*.

4.3.3 Les limites d'une machine dans la détermination de la causalité exprimée dans un document

Concernant l'apprentissage automatique pour la classification des phrases causales, nous pensons pour le moment qu'une limite hypothétique à notre travail est la nécessité d'utiliser des documents d'apprentissage qui traitent sensiblement du même sujet (*topic*) que les documents à classifier. Il faut comparer ce qui est comparable et les ressources allouées étant limitées (en temps, en capacités, en compétences) nous gageons que l'efficacité de notre machine sera d'autant meilleure si on lui donne à classifier des documents comparables au document d'apprentissage. Mais à terme, nous pouvons espérer soit de la nullité de cette limite, soit de nous en affranchir.

Une deuxième limite est le choix de l'algorithme de classification (supervisé par construction) : nous avons porté notre dévolu sur des bayésiens naïfs simples que nous faisons travailler en réseau, assisté éventuellement par un algorithme de similarité de surface. Nous avons choisi ces algorithmes car leur fonctionnement est explicable et leur « raisonnement » prévisible. C'est un choix pragmatique, mais l'arrivée de nouveaux algorithmes d'apprentissage profonds, comme les réseaux de neurones, dont le fonctionnement est plus ardu à expliquer, pourraient être envisagés par la suite. La

proposition majeure de cette thèse souffre également de quelques limites. La première limite est la création de référentiels pour l'apprentissage et construction de règles d'inférences :

- le référentiel *Csas* pour l'algorithme *NER ;*
    - la couche syntaxico-sémantique ;
- la coréférence, grammaticale et lexicale ;
    - créer un dictionnaire (limité au domaine de définition) d'unités considérées comme synonymes ;
- la classification manuelle des phrases causales, la création du document d'apprentissage ;
- la base de données de règles de forage des arguments *Cause* et *Effect*.

Notre machine travaille sur la sémantique et la syntaxe d'une phrase ; une phrase à la fois, dans l'ensemble du document. L'échelle minimale d'analyse est le morphème. Nous restons donc dans l'analyse structurelle logique, c'est-à-dire directement déduite de la grammaire et de son formalisme. On pourrait penser maintenant, compte tenu de l'avancée en matière d'apprentissage automatique, supervisé et non-supervisé, que l'on pourrait obtenir de meilleurs résultats, non pas en s'affranchissant totalement de ce parallélisme (cela n'aurait aucun sens, les choses doivent rester explicables), mais en le renforçant à l'aide d'algorithmes qui « regarderaient » d'une autre manière la phrase et ses constituants, ou encore la phrase dans le document. Concernant l'apprentissage non-supervisé (non traité dans cette thèse car nous avons un objectif défini de classification), il pourrait y avoir des hypothèses à vérifier qui permettrait d'aider à la classification des phrases d'évènements ou causales (par définition, il n'est pas possible de savoir à l'avance si ces hypothèses ont un intérêt en rapport à notre problème, c'est à l'algorithme de le déterminer). Après une classification effectuée par un algorithme d'apprentissage supervisé, dans le but de renforcer le modèle, on pourrait s'intéresser par exemple à :

- la longueur des phrases (le nombre de « mots ») ;
- leur position absolue ou relative (des phrases) dans un document.

Cependant, nous croyons beaucoup à la capacité de l'empreinte syntaxique que nous avons façonnée dans le chapitre 3 pour aider à l'élucidation des phrases causales, en ce qu'elle permettrait de se détacher quelque peu de l'empreinte sémantique. Il y a un enjeu important à relever sur ce point-là.

Une limite importante à notre machine, à l'état de théorie, est le corollaire de ce que nous avons exposé plus tôt à propos de la causalité contrefactuelle. Notre machine n'est pas capable pour le moment de faire la différence entre une causalité affirmée, déclarée comme telle, et une causalité contrefactuelle, une hypothèse d'explication causale sur le monde. En phase de détection, nous pouvons *a priori* penser que nous serons virtuellement capables de cerner toutes les expressions de causalité existantes

dans une langue[248] : c'est cependant un avantage certain. Mais il faudra être capable de proposer un moyen de discernement entre ce qui relève de la déclaration, de l'affirmation et de ce qui est contraire aux faits ou hypothétique.

248 Il y aura un travail conséquent de réflexion à mettre en œuvre en amont. Il nécessite la rencontre entre experts du domaine (linguistique, grammaire anglaise, informatique, *machine learning*) et la mise en place d'une équipe pour effectuer un travail de classification et d'annotation de documents qui serviront de modèles pour l'apprentissage de notre machine. C'est un projet à mettre en œuvre.

Conclusion

Cette thèse a eu pour objectif de présenter le forage des données et la formalisation des connaissances sur un accident en s'appuyant sur le cas *Deepwater Horizon*. Elle s'inscrit à la croisée des chemins entre les humanités, *safety studies* et les sciences de l'ingénieur. Notre démarche d'expert qui a guidé notre réflexion et l'appui théorique apporté par le cadre conceptuel de l'ingénierie en situation extrême nous ont permis de dépasser l'approche strictement technique et d'aller explorer « hors des sentiers battus » ce cas d'accident. L'étude de ce que nous avons appelé l'intervention a été le point focal de nos travaux, source de nos apports et de nos premiers résultats. Cette période de l'accident, qui dura quatre-vingt-sept jours, a été très largement sous-estimée par la science, restant dans des considérations « classiques » de l'accident, les explosions et la « marée noire ». Nous avons été peu ou prou jeté dans le grand bain face à un cas qui méritait que l'on s'y penche en profondeur. De là, est née notre réflexion sur le forage des données, dont nous tentons une définition ajustée par l'expérience de ces travaux de thèse. D'abord considérer qu'il n'est plus objectivement possible de travailler sur la littérature d'un évènement, quel qu'il soit, sans assistance informatique (algorithmique) pour traiter le texte en question, compte tenu du volume disponible de données, de l'hétérogénéité des sources et des impératifs d'exigences de qualité et de pertinence. Ensuite, à défaut de pouvoir modéliser, non pas par difficulté, mais faute de modèle existant, il s'agit au moins de pouvoir représenter par ses connaissances l'objet étudié : l'utilisation des ontologies de haut niveau doit être encouragée et le processus de population doit être automatisé. L'« atome de connaissance » ainsi créé (le *triple*), doit devenir un point de passage obligé en matière de représentation des connaissances, car il est irréductible, donc explicite, et parfaitement compréhensible par un humain comme par une machine. Enfin, encourager la création d'algorithmes d'apprentissage automatique pour l'analyse textuelle en masse et la possibilité de faire émerger des connaissances nouvelles à partir des résultats obtenus.

Nous l'avons vu, la connaissance, particulièrement scientifique, était lacunaire sur l'intervention et ses quatre-vingt-sept jours de lutte, l'utilisation de modèles dépassés y étaient probablement pour quelque chose et l'on voit qu'une redéfinition de « l'accident », au prisme de l'ingénierie en situation extrême nous a permis d'apporter non pas un regard nouveau, mais bien un regard premier sur cet accident. Peu de chose avait été écrite en français auparavant à propos de l'accident de *Deepwater Horizon*. Et comme nous l'avons montré, la littérature en anglais est certes beaucoup plus conséquente, mais pas forcément plus robuste pour autant. Les connaissances que nous avons produites devraient pouvoir servir de point de départ pour la continuité des recherches sur ce cas. Il y a de belles perspectives de recherche qui s'ouvrent, tant

sur l'accident en tant que tel que sur les données qu'il a produites. Nous pensons cependant que nous avons comblé, au moins en langue française, un manque important en matière de connaissances à propos de cet accident.

Nous avons abordé l'accident de *Deepwater Horizon* particulièrement *via* les solutions d'ingénierie déployées pour résoudre un problème sans précédent. Nous avons majoritairement exploité les rapports d'enquêtes et les quelques papiers de science, mais n'avons pas assez exploité les trois entretiens que nous avons menés en avril 2016. Nous avons accédé à un degré de finesse sans précédent à la connaissance de cet accident par ces témoignages de l'intérieur, mais il nous paraît possible, et même nécessaire de les exploiter bien au-delà du spectre « pur » de l'ingénierie et de la causalité évènementielle comme nous l'avons fait. Ils doivent pouvoir servir de matériau pour des recherches ultérieures.

Si nous élargissons le spectre des données à la recherche d'informations en source ouverte, en montrant le traitement famélique de cet accident par la science, nous avons aussi montré l'indispensable intérêt de trouver des « artefacts algorithmiques » pour assister le chercheur dans sa revue de littérature. Nous avons soulevé ce problème d'épistémologie par le fait qu'il est foncièrement impossible pour un humain de lire les quelques milliers d'articles scientifiques qui traitent peu ou prou de même sujet d'étude. Notre embryon de démarche dans le cadre du forage des données, présentée dans le chapitre 1, s'apparente à un rasoir d'Ockham, qui permet *a minima* de pouvoir se justifier de la mise au rebut d'une quantité non négligeable d'informations finalement peu pertinentes eut égard au sujet en question. Ce rasoir doit aujourd'hui être algorithmique.

Puis nous avons exploré en profondeur les ontologies, en particulier *DOLCE*, pour, d'une part, apporter un formalisme indispensable à la connaissance, et d'autre part construire un référentiel conceptuel pour l'algorithme *NER* que nous avons pensé afin de résoudre, en partie, le problème conséquent qu'est le processus de population d'une ontologie. Notre ontologie de l'accident de *Deepwater Horizon* doit pouvoir servir de matériau de recherche. Il faudra dans l'avenir, à la manière de ce qui se fait déjà pour les taxinomies biologiques (qui ne sont pas des ontologies, mais qui s'en rapprochent le plus), et dans une moindre mesure par *DBpedia*, proposer une interface utilisateur qui permette aisément d'explorer l'ontologie. Le projet le plus immédiat à mettre en œuvre est le développement informatique de l'algorithme *NER* que nous avons proposé ; si les résultats sont bons, nous aurons apporté une réelle amélioration opérationnelle en ingénierie des connaissances. A plus court terme, il y a une piste intéressante que nous avons succinctement présentée dans le chapitre 2 qui est le « récit sémantique ». Un objectif intéressant à poursuivre pourrait être, par l'utilisation de logiciels de gestion de contenu comme *MediaWiki,* associés à un moteur sémantique comme *Semantic MediaWiki*, de proposer un récit sémantique où *DOLCE* serait l'ontologie qui structurerait et formaliserait le récit. Ainsi, on mettrait à disposition de la communauté une « base de connaissances » et au-delà, on pourrait proposer ce nouveau modèle d'ingénierie des connaissances comme une référence. De la sorte, au moins pour l'utilisateur final, on s'affranchirait du problème de visualisation et d'exploration de l'ontologie, devenue fondation du récit.

Enfin, nous interrogeons la causalité de différentes manières. Pour l'aborder, nous avons d'abord travaillé sur les solutions d'ingénierie déployées et particulièrement sur l'échec du *cofferdam*. Nous avons d'abord montré l'intérêt de l'*ESA* et du logiciel *ETHNO* pour l'analyse de la causalité dans les récits, mais nous avons aussi trouvé des limites qui nous ont poussés à aller voir au-delà et nous aventurer sur le terrain très actuel de l'apprentissage automatique pour apporter une réponse opérationnelle aux limites actuelles en matière d'analyse de la causalité exprimée dans les textes. Nous avons donc proposé une méthode algorithmique d'apprentissage supervisé pour être capable de cerner ces expressions et construire le « cheminement causal » dans un document écrit. En rappelant les enjeux liés à l'interprétation des rapports d'enquêtes en matière de prescriptions et de recommandations réglementaires, nous pensons que donner l'accès à la compréhension et à l'évaluation de l'enquête et du raisonnement des enquêteurs est une avancée importante. Notre machine IA sera à même de répondre à la question *pourquoi ?* qu'un lecteur humain pourra se poser.

Conscient de la prévalence du raisonnement contrefactuel dans l'expression des causes d'un évènement, mais de son empreinte textuelle *a priori* moins marquée que les autres raisonnements causaux, le prolongement immédiat que nous voyons à nos travaux d'algorithmie est la mise en place d'une classification donnant la possibilité de déterminer si la causalité exprimée est avérée ou supposée. Pour le moment, nous avons proposé une classification à « un étage », la phrase est causale ou ne l'est pas, et nous pensons qu'il sera très vite intéressant de pouvoir qualifier la causalité exprimée comme une affirmation ou une hypothèse contrefactuelle. Et ce, encore une fois, détachée de toute volonté de quête de vérité, simplement pour cerner les expressions de causalité où le narrateur exprime une affirmation, une déclaration, une démonstration de factualité pour appuyer la causalité, ou bien si, au contraire, il émet des doutes, des conditions, des hypothèses ou encore des réserves à ce qu'il exprime, ou qu'il « refait le monde » pour tenter de comprendre la survenance d'un évènement. Cette classification « à deux étages » pourra se faire exactement sur les mêmes bases que la première classification, mais il faudra d'autant plus de travail d'annotation puisqu'il s'effectuera *a priori* exclusivement sur les phrases causales. Encore une fois, nous pensons que l'empreinte syntaxique pourra être une clé très efficace, en sus de l'empreinte sémantique, pour la détermination d'une phrase à la causalité avérée ou supposée. Le premier travail à effectuer à ce sujet est de cerner l'expression du doute dans la langue étudiée. On voit également que, poussé à l'extrême, on ira explorer la sémantique de l'accident dans ses dimensions de croyance ou de mystique ; nous l'avons montré, même les rapports d'enquêtes ne sont pas exempts d'une grande part d'« irrationnel », dans leur production et dans leur contenu.

Par la suite, une fois que notre algorithme aura appris convenablement à identifier ces phrases, on pourra proposer des cheminements causaux avec une surcouche « qualité de la causalité » qui mettra en exergue les zones de doutes dans l'agencement causal des évènements. En poussant encore plus loin, on pourrait arriver à mettre en parallèle le cheminement causal factuel et le cheminement contrefactuel exprimés tous les deux dans le même document à propos du même évènement.

A plus long terme, cette recherche peut s'inscrire dans un projet plus ambitieux portant sur la possibilité de créer une machine IA qui pourrait évaluer la qualité de l'argumentation causale exprimée dans n'importe quel type de texte. Cette machine pourrait aider à rédiger, avec un forage de données encore plus robuste, la structure d'un rapport d'enquête « en temps réel ». Lorsqu'un accident survient et qu'une équipe d'enquêteurs est à l'ouvrage, cette machine pourrait les assister dans leur rédaction en évaluant leur réflexion et en leur proposant des « solutions causales » robustes en matière de raisonnement et d'arguments. Nous pouvons espérer que cela pourrait améliorer l'argumentation en profondeur et aider à faire émerger des causes au préalable plus difficiles à déterminer.

Bibliographie

Accidents, événementialité et causalité

Because, because of and cos, cos of - English Grammar Today - Cambridge Dictionary (no date). Available at: https://dictionary.cambridge.org/grammar/british-grammar/because-because-of-and-cos-cos-of (Accessed: 19 February 2018).

Benner Jr., L. (1975) 'Accident Investigations: Multilinear Events Sequencing Methods', *Journal of Safety Research*, 7(2), pp. 67–73.

Benner Jr., L. (1985) 'Rating Accident Models and Investigation Methodologies', *Journal of Safety Research*, 16, pp. 105–126.

Benner Jr., L. (1989) 'The MES Investigator's Handbook'. Starline Software Ltd. Available at: http://www.ludwigbenner.org/arch6.html.

Burns, C. P. (2000) *Analysing accident reports using structured and formal methods*. PhD. University of Glasgow. Available at: http://encore.lib.gla.ac.uk/iii/encore/record/C__Rb1891139 (Accessed: 22 July 2017).

Callon, M. (1986) 'Eléments pour une sociologie de la traduction', *L'Année sociologique*. Edited by M. Akrich, M. Callon, and B. Latour, (36), pp. 169–208.

CAUSALITÉ : Définition de CAUSALITÉ (no date). Available at: http://www.cnrtl.fr/definition/causalit%C3%A9 (Accessed: 11 October 2018).

COMPRENDRE : Définition de COMPRENDRE (no date). Available at: http://www.cnrtl.fr/definition/comprendre (Accessed: 26 January 2018).

Emerson, R. M. (1981) 'On Last Resorts', *American Journal of Sociology*, 87(1), pp. 1–22. doi: 10.1086/227417.

Event Structure Analysis (no date). Available at: http://www.indiana.edu/%7Esocpsy/papers/modelingEvents/esa.htm (Accessed: 15 October 2018).

Gephart, R. P. (1984) 'Making Sense of Organizationally Based Environmental Disasters', *Journal of Management*, 10(2), pp. 205–225. doi: 10.1177/014920638401000205.

Gephart, R. P. (1992) 'Sensemaking, communicative distortion and the logic of public inquiry legitimation', *Organization & Environment*, 6(2), pp. 115–135.

Gephart, R. P. (1993) 'The Textual Approach: Risk and Blame in Disaster Sensemaking', *Academy of Management Journal*, 36(6), pp. 1465–1514. doi: 10.2307/256819.

Gephart, R. P. (1997) 'Hazardous measures: an interpretive textual analysis of quantitative sensemaking during crises', *Journal of Organizational Behavior*, 18(S1), pp. 583–622.

Gephart, R. P., Steier, L. and Lawrence, T. (1990) 'Cultural rationalities in crisis sensemaking: a study of a public inquiry into a major industrial accident', *Organization & Environment*, 4(1), pp. 27–48.

Gephart, R. P., Topal, C. and Zhang, Z. (2010) 'Future-Oriented Sensemaking: Temporalities and Institutional Legitimation', in Hernes, T. and Maitlis, S. (eds) *Process, Sensemaking & Organizing*. New York NY: Oxford University Press (Perspectives on Process Organization Studies), pp. 275–302.

Girju, R. (2003) 'Automatic detection of causal relations for question answering', in *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*. Association for Computational Linguistics, pp. 76–83.

Girju, R. and Moldovan, D. I. (2002) 'Text mining for causal relations.', in *FLAIRS conference*, pp. 360–364.

Griffin, L. J. (1993) 'Narrative, Event-Structure Analysis, and Causal Interpretation in Historical Sociology', *The American Journal of Sociology*, 98(5), pp. 1094–1133.

Griffin, L. J. (2007) 'Historical Sociology, Narrative and Event-Structure Analysis: Fifteen Years Later', *Sociologica*, (3), pp. 0–0. doi: 10.2383/25956.

Griffin, L. J. and Korstad, R. R. (1998) 'Historical Inference and Event-Structure Analysis', *International Review of Social History*, 43(S6), pp. 145–165. doi: 10.1017/S0020859000115135.

Grivaz, C. (2010) 'Human Judgements on Causation in French Texts', in Chair), N. C. (Conference et al. (eds) *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).

Guarnieri, F. and Travadel, S. (2018) *Un récit de Fukushima*. Paris: PUF.

Heise, D. R. (no date a) 'Event Structure Analysis with Ethno', p. 29.

Heise, D. R. (no date b) 'Event Structure Analysis with Ethno', p. 29.

Investigation Quality Assurance (no date). Available at: http://www.iprr.org/lib/QMA_P1.html (Accessed: 22 July 2017).

Koen, B. V. (1985) *Definition of the engineering method*. American Society for Engineering Education. Washington, DC.

Laukkanen, M. (2008) 'Comparative Causal Mapping with CMAP3', *A Method Introduction to Comparative Causal Mapping and a User's Manual for CMAP3. Kuopio: Kuopio University Occasional Reports H. Business and Information Technology*, 2.

Laukkanen, M. (2012) 'Comparative Causal Mapping and CMAP3 Software in Qualitative Studies', *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 13(2). Available at: http://www.qualitative-research.net/index.php/fqs/article/view/1846 (Accessed: 29 December 2017).

Leveson, N. G. (2011) *Engineering a Safer World*. Cambridge, MA; London: The MIT Press (Engineering Systems). Available at: https://mitpress.mit.edu/sites/default/files/titles/free_download/9780262016629_Engineering_a_Safer_World.pdf (Accessed: 14 January 2016).

PPRT (no date) *Site national PPRT - Inspection des Installations Classées*. Available at: http://www.installationsclassees.developpement-durable.gouv.fr/-Site-national-PPRT-.html (Accessed: 4 July 2017).

Qureshi, Z. H. (2008) *A review of accident modelling approaches for complex critical sociotechnical systems*. Technical Report. Available at: http://dspace.dsto.defence.gov.au/dspace/handle/1947/9120 (Accessed: 26 May 2017).

Reason, J. (2013) *L'erreur humaine*. Translated by J.-M. Hoc. Paris: Presses des Mines.

Sklet, S. (2004) 'Comparison of some selected methods for accident investigation', *Journal of Hazardous Materials*, 111(1–3), pp. 29–37. doi: 10.1016/j.jhazmat.2004.02.005.

The ESA Site (no date). Available at: http://www.indiana.edu/~socpsy/ESA/ (Accessed: 17 February 2018).

Travadel, S. and Guarnieri, F. (2015) 'L'agir en situation extrême', in Guarnieri, F. et al. (eds) *L'accident de Fukushima Dai Ichi, Le récit du directeur de la centrale*. Paris: Presses des Mines, pp. 283–321.

Travadel, S. and Guarnieri, F. (2016) 'Modèles d'accident et temporalités', in Guarnieri, F. et al. (eds) *L'accident de Fukushima Dai Ichi*. Paris: Presses des Mines, pp. 23–42.

Travadel, S. and Guarnieri, F. (2018) 'Dessine-moi un accident…', *Préventique*, (159), pp. 36–37.

Travadel, S., Guarnieri, F. and Portelli, A. (2018) 'Industrial Safety and Utopia: Insights from the Fukushima Daiichi Accident', *Risk Analysis*, 38(1), pp. 56–70.

Wilde, G. J. S. (1982) 'The Theory of Risk Homeostasis: Implications for Safety and Health', *Risk Analysis*, 2(4), pp. 209–225. doi: 10.1111/j.1539-6924.1982.tb01384.x.

A Deeper Look into the Java 8 Date and Time API - DZone Java (no date) *dzone.com*. Available at: https://dzone.com/articles/deeper-look-java-8-date-and (Accessed: 20 May 2018).

Abeillé, A., Clément, L. and Toussenel, F. (2003) 'Annotation morpho-syntaxique', *Paper available at http://www. llf. cnrs. fr/Gens/Abeille/guide-morphosynt*, 2.

Abeillé, A., Clément, L. and Toussenel, François (2003) 'Building a treebank for French', in *Treebanks*. Springer, pp. 165–187.

Abeillé, A., Toussenel, F. and Chéradame, M. (2004) 'Corpus le monde: Annotations en constituants. guide pour les correcteurs', *LLF, UFRL, Paris7*.

Algorithm::NaiveBayes - search.cpan.org (no date). Available at: http://search.cpan.org/~kwilliams/Algorithm-NaiveBayes-0.04/lib/Algorithm/NaiveBayes.pm (Accessed: 29 April 2018).

AllenNLP - Models (no date). Available at: http://allennlp.org/models (Accessed: 14 May 2018).

Arlot, S. (2017) 'Validation croisée', p. 37.

Asghar, N. (2016) 'Automatic Extraction of Causal Relations from Natural Language Texts: A Comprehensive Survey', *arXiv:1605.07895 [cs]*. Available at: http://arxiv.org/abs/1605.07895 (Accessed: 13 February 2018).

Bellinger, G., Castro, D. and Mills, A. (2004) 'Data, information, knowledge, and wisdom'.

Bernstein, J. H. (2011) 'The data-information-knowledge-wisdom hierarchy and its antithesis', *Nasko*, 2(1), pp. 68–75.

Bidirectional Attention Flow for Machine Comprehension - Semantic Scholar (no date). Available at: /paper/Bidirectional-Attention-Flow-for-Machine-Seo-Kembhavi/007ab5528b3bd310a80d553cccad4b78dc496b02 (Accessed: 13 May 2018).

Bill Buchanan (no date) *Napier*. Available at: https://www.napier.ac.uk:443/people/bill-buchanan/ (Accessed: 20 May 2018).

Chen, D. and Manning, C. (2014) 'A fast and accurate dependency parser using neural networks', in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 740–750.

De Marneffe, M.-C. *et al.* (2014) 'Universal Stanford dependencies: A cross-linguistic typology.', in *LREC*, pp. 4585–4592.

De Marneffe, M.-C. and Manning, C. D. (2008) *Stanford typed dependencies manual*. Technical report, Stanford University.

Domingos, P. (2012) 'A few useful things to know about machine learning', *Communications of the ACM*, 55(10), p. 78. doi: 10.1145/2347736.2347755.

Domingos, P. and Pazzani, M. (1997) 'On the Optimality of the Simple Bayesian Classifier under Zero-One Loss', *Machine Learning*, 29(2–3), pp. 103–130. doi: 10.1023/A:1007413511361.

Drools - Drools - Business Rules Management System (Java™, Open Source) (no date). Available at: https://www.drools.org/ (Accessed: 4 August 2018).

Duby, C. and Robin, S. (2006) 'Analyse en composantes principales', *Institut National Agronomique, Paris-Grignon*, 80.

Finkel, J. R., Grenager, T. and Manning, C. (2005) 'Incorporating non-local information into information extraction systems by gibbs sampling', in *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, pp. 363–370.

Gardner, M. *et al.* (2018) 'AllenNLP: A Deep Semantic Natural Language Processing Platform', *arXiv:1803.07640 [cs]*. Available at: http://arxiv.org/abs/1803.07640 (Accessed: 13 May 2018).

Glossaire du machine learning (no date) *Google Developers*. Available at: https://developers.google.com/machine-learning/glossary/?hl=fr (Accessed: 4 November 2018).

H.Gomaa, W. and A. Fahmy, A. (2013) 'A Survey of Text Similarity Approaches', *International Journal of Computer Applications*, 68(13), pp. 13–18. doi: 10.5120/11638-7118.

Higashinaka, R. and Isozaki, H. (2008) *Automatically Acquiring Causal Expression Patterns from Relation-annotated Corpora to Improve Question Answering for why-Questions*.

HIGUCHI, Koichi (no date). Available at: http://koichi.nihon.to/psnl/en/ (Accessed: 28 September 2017).

'Introduction au TALN et à l'ingénierie linguistiqu.pdf' (no date). Available at: http://www.lattice.cnrs.fr/sites/itellier/poly_info_ling/info-ling.pdf (Accessed: 1 February 2018).

java - How to sort timestamp list? (no date) *Stack Overflow*. Available at: https://stackoverflow.com/questions/44238533/how-to-sort-timestamp-list (Accessed: 20 May 2018).

java - Sort objects in ArrayList by date? (no date) *Stack Overflow*. Available at: https://stackoverflow.com/questions/5927109/sort-objects-in-arraylist-by-date (Accessed: 20 May 2018).

JoliCode (no date) *Xavier Lacot*, *JoliCode*. Available at: https://jolicode.com/equipe/xavier-lacot (Accessed: 28 January 2018).

Kahlmann, A. (1975) *Traitement automatique d'un dictionnaire de synonymes. Etude de sa structure. Méthode de contrôle et de perfectionnement*. Stockholm.

KH Coder / Discussion / Open Discussion:Co-occurrence Networks (no date a). Available at: https://sourceforge.net/p/khc/discussion/222396/thread/2da0ff02/ (Accessed: 28 September 2017).

KH Coder / Discussion / Open Discussion:Co-occurrence Networks (no date b). Available at: https://sourceforge.net/p/khc/discussion/222396/thread/2da0ff02/ (Accessed: 28 September 2017).

Koichi HIGUCHI (2016) *KH Coder 3 Reference Manual*. Available at: http://khc.sourceforge.net/en/manual_en_v3.pdf (Accessed: 28 September 2017).

de La Clergerie, É. V. *et al.* (2008) 'Large scale production of syntactic annotations for french', in *First Workshop on Automated Syntactic Annotations for Interoperable Language Resources*.

Manning, C. *et al.* (2014) 'The Stanford CoreNLP natural language processing toolkit', in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55–60.

McCallum, A. and Nigam, K. (1998) 'A comparison of event models for Naive Bayes text classification', in *In Aaai-98 Workshop on Learning for Text Categorization*. AAAI Press, pp. 41–48.

Rajpurkar, P. *et al.* (2016) 'SQuAD: 100,000+ Questions for Machine Comprehension of Text', *arXiv:1606.05250 [cs]*. Available at: http://arxiv.org/abs/1606.05250 (Accessed: 13 May 2018).

Rink, B., Bejan, C. A. and Harabagiu, S. M. (2010) 'Learning Textual Graph Patterns to Detect Causal Event Relations.', in *FLAIRS Conference*.

Sasaki, Y. (2007) 'The truth of the F-measure', p. 5.

Schmid, H. (1994) *Probabilistic Part-of-Speech Tagging Using Decision Trees*.

Schmid, H. (1995) 'Improvements In Part-of-Speech Tagging With an Application To German', in *In Proceedings of the ACL SIGDAT-Workshop*, pp. 47–50.

Seo, M. J. *et al.* (2016) 'Bidirectional Attention Flow for Machine Comprehension', *CoRR*, abs/1611.01603.

Silva, M. A. G. (2018) *SimMetrics is a Similarity Metric Library for strings (as of http://sourceforge.net/projects/simmetrics/)*. Available at: https://github.com/magsilva/SimMetrics (Accessed: 21 May 2018).

Smith, T. F. and Waterman, M. S. (1981) 'Identification of common molecular subsequences', *Journal of Molecular Biology*, 147(1), pp. 195–197. doi: 10.1016/0022-2836(81)90087-5.

Stanford CoreNLP – Natural language software | Stanford CoreNLP (no date). Available at: https://stanfordnlp.github.io/CoreNLP/index.html (Accessed: 18 March 2018).

Stanford Knowledge Systems, AI Laboratory (no date). Available at: http://ksl.stanford.edu/ (Accessed: 25 January 2018).

Teaching - Smith-Waterman (no date). Available at: http://rna.informatik.uni-freiburg.de/Teaching/index.jsp?toolName=Smith-Waterman (Accessed: 21 May 2018).

Tellier, I. (no date) *Introduction au TALN et à l'ingénierie linguistique, université*. Available at: http://www.lattice.cnrs.fr/sites/itellier/poly_info_ling/index.html (Accessed: 31 January 2018).

Teulier, R., Charlet, J. and Tchounikine, P. (2005) 'L'ingénierie des connaissances: acquis et nouvelles perspectives', in *Ingénierie des connaissances*. L'Harmattan, pp. 11–26.

The Berkeley NLP Group (no date). Available at: http://*NLP*.cs.berkeley.edu/software.shtml (Accessed: 5 February 2018).

The Stanford Natural Language Processing Group (no date). Available at: https://*NLP*.stanford.edu/ (Accessed: 28 December 2017).

The Stanford Question Answering Dataset (no date). Available at: https://rajpurkar.github.io/SQuAD-explorer/ (Accessed: 28 April 2018).

Théorème de Bayes (no date). Available at: http://statelem.com/theoreme_de_bayes.php (Accessed: 7 May 2018).

Vogler, R. (2013) 'Comparison of String Distance Algorithms', *joy of data*, 21 August. Available at: https://www.joyofdata.de/blog/comparison-of-string-distance-algorithms/ (Accessed: 24 May 2018).

What is text similarity? | Kavita Ganesan (no date). Available at: http://kavita-ganesan.com/what-is-text-similarity/ (Accessed: 20 May 2018).

Ackoff, R. L. (1989) 'From data to wisdom', *Journal of applied systems analysis*, 16(1), pp. 3–9.

Alani, H. *et al.* (2003) 'Automatic ontology-based knowledge extraction from Web documents', *IEEE Intelligent Systems*, 18(1), pp. 14–21. doi: 10.1109/MIS.2003.1179189.

Amardeilh, F. (2006) 'OntoPop or how to annotate documents and populate ontologies from texts', in *ESWC 2006 Workshop on Mastering the Gap: From Information Extraction to Semantic Representation*. CEUR Workshop Proceedings.

Amsili, P. (2006) 'Logique du premier ordre (Une introduction pour les linguistes)'. Université de Paris 7, Atelier logique et s´emantique du langage naturel Semaine bordelaise de sémantique formelle.

AXIOME : Définition de AXIOME (no date). Available at: http://www.cnrtl.fr/definition/axiome (Accessed: 4 January 2018).

Baader, F., Horrocks, I. and Sattler, U. (2008) 'Chapter 3 Description Logics', in *Foundations of Artificial Intelligence*. Elsevier, pp. 135–179. doi: 10.1016/S1574-6526(07)03003-9.

Babitski, G. *et al.* (2009) 'Ontology Design for Information Integration in Disaster Management.', *GI Jahrestagung*, 154, pp. 3120–3134.

Barry Smith (no date). Available at: http://www.buffalo.edu/cas/philosophy/faculty/faculty_directory/smith-b.html (Accessed: 25 January 2018).

Bates, M. J. (1989) 'The design of browsing and berrypicking techniques for the online search interface', *Online review*, 13(5), pp. 407–424.

Batres, R. *et al.* (2007) 'An upper ontology based on ISO 15926', *Computers & Chemical Engineering*. (ESCAPE-15), 31(5), pp. 519–534. doi: 10.1016/j.compchemeng.2006.07.004.

Batres, R. *et al.* (2014) 'The use of ontologies for enhancing the use of accident information', *Process Safety and Environmental Protection*, 92(2), pp. 119–130. doi: 10.1016/j.psep.2012.11.002.

Bénel, A. (2011) 'D'où viennent les ontologies? À la recherche du chaînon manquant', in *Atelier Philosophie et Ingénierie. Le formel face à l'histoire, la technologie et la matérialité (IC2011)*, p. 2.

Borst, W. N. (1997) *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. Available at: https://research.utwente.nl/en/publications/construction-of-engineering-ontologies-for-knowledge-sharing-and- (Accessed: 25 January 2018).

Chapuis-Schmitz, D. (2004) 'Le cercle de Vienne', *Labyrinthe*, (18), pp. 11–16. doi: 10.4000/labyrinthe.200.

Charlet, J. (2002) 'L'ingénierie des connaissances: développements, résultats et perspectives pour la gestion des connaissances médicales'.

Chez Laurent Roussarie (no date). Available at: http://l.roussarie.free.fr/ (Accessed: 25 January 2018).

CHOIR in Clinical Practice (no date) 'CHOIR in Clinical Practice'. Available at: https://choir.stanford.edu/clinical-practice/ (Accessed: 18 July 2017).

Cmap | Cmap Software (no date). Available at: https://cmap.ihmc.us/publications/research-publications.php (Accessed: 28 January 2018).

CmapTools Ontology Editor (no date). Available at: http://cmap.ihmc.us/coe/test/v401ReleaseNotes.html (Accessed: 22 January 2018).

Davalcu, H. *et al.* (2003) 'OntoMiner: bootstrapping and populating ontologies from domain-specific Web sites', *IEEE Intelligent Systems*, 18(5), pp. 24–33. doi: 10.1109/MIS.2003.1234766.

David Ellis (1984) 'Theory and explanation in information retrieval research', *Information Scientist*, 8(1), pp. 25–38. doi: 10.1177/016555158400800105.

David Ellis (1989) 'A behavioural model for information retrieval system design', *Journal of Information Science*, 15(4–5), pp. 237–247. doi: 10.1177/016555158901500406.

DBpedia (no date). Available at: http://wiki.dbpedia.org/ (Accessed: 28 January 2018).

DBpedia Wiki | DBpedia (no date). Available at: http://wiki.dbpedia.org/dbpedia-wiki (Accessed: 28 January 2018).

Delir Haghighi, P. *et al.* (2013) 'Development and evaluation of ontology for intelligent decision support in medical emergency management for mass gatherings', *Decision Support Systems*, 54(2), pp. 1192–1204. doi: 10.1016/j.dss.2012.11.013.

Draghici, S. *et al.* (2003) 'Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate', *Nucleic Acids Research*, 31(13), pp. 3775–3781.

Dumez, H. (2016) *Comprehensive Research: A Methodological and Epistemological Introduction to Qualitative Research*. Copenhagen: Copenhagen Business School Press.

Dumez, H. and Rigaud, E. (2008) 'Comment passer du matériau de recherche à l'analyse théorique? A propos de la notion de template', *Le Libellio d'Aegis*, 4(2), pp. 40–46.

Eder, C. *et al.* (2005) 'Pragmatic strategies that enhance the reliability of data abstracted from medical records', *Applied Nursing Research*, 18(1), pp. 50–54. doi: 10.1016/j.apnr.2004.04.005.

'FaCT++ reasoner | *OWL* research at the University of Manchester' (no date). Available at: http://*owl*.cs.manchester.ac.uk/tools/fact/ (Accessed: 28 January 2018).

Fiorentini, X. *et al.* (2013) 'Modeling nuclear power plants engineering data using ISO 15926', in *Proceedings of 2013 International Conference on Industrial Engineering and Systems Management (IESM). Proceedings of 2013 International Conference on Industrial Engineering and Systems Management (IESM)*, pp. 1–6.

FORMEL : Définition de FORMEL (no date). Available at: http://www.cnrtl.fr/definition/formel (Accessed: 25 January 2018).

Francis, J. M. *et al.* (1999) 'Performance Measures For Information Extraction', in *In Proceedings of DARPA Broadcast News Workshop*, pp. 249–252.

Frost, M. H. *et al.* (2007) 'What Is Sufficient Evidence for the Reliability and Validity of Patient-Reported Outcome Measures?', *Value in Health*, 10, pp. S94–S105. doi: 10.1111/j.1524-4733.2007.00272.x.

Gangemi, A. *et al.* (2002) 'Sweetening ontologies with DOLCE', in *Knowledge engineering and knowledge management: Ontologies and the semantic Web*. Springer, pp. 166–181. Available at: http://link.springer.com/chapter/10.1007/3-540-45810-7_18 (Accessed: 12 February 2016).

Gangemi, A. *et al.* (2003) 'Sweetening WORDNET with DOLCE', *AI Magazine*, 24(3), p. 13. doi: 10.1609/aimag.v24i3.1715.

Gangemi, A. *et al.* (2013) 'Fred as an event extraction tool', in *Proceedings of the Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web*, p. 14.

Gangemi, A. *et al.* (2016) 'Semantic web machine reading with FRED', *Semantic Web*, (Preprint), pp. 1–21.

Geleijnse, G. and Korst, J. H. (2005) 'Automatic Ontology Population by Googling.', in *BNAIC*, pp. 120–126.

Genesereth, M. R. and Nilsson, N. J. (1987) 'Logical foundations of artificial', *Intelligence. Morgan Kaufmann*, 2.

Gilbert, E. H. *et al.* (1996) 'Chart Reviews In Emergency Medicine Research: Where Are The Methods?', *Annals of Emergency Medicine*, 27(3), pp. 305–308. doi: 10.1016/S0196-0644(96)70264-0.

Gregory, K. E. and Radovinsky, L. (2012) 'Research strategies that result in optimal data collection from the patient medical record', *Applied Nursing Research*, 25(2), pp. 108–116. doi: 10.1016/j.apnr.2010.02.004.

Gruber, T. R. (1993a) 'A translation approach to portable ontology specifications', *Knowledge acquisition*, 5(2), pp. 199–220.

Gruber, T. R. (1993b) 'What is an Ontology', *WWW Site http://www-ksl. stanford. edu/kst/whatis-an-ontology. html (accessed on 07-09-2004)*. Available at: http://ejournal.narotama.ac.id/files/Ontology..pdf (Accessed: 15 February 2016).

Gruber, T. R. (1995) 'Toward principles for the design of ontologies used for knowledge sharing?', *International journal of human-computer studies*, 43(5–6), pp. 907–928.

Guarino, N. (1995) 'Formal ontology, conceptual analysis and knowledge representation', *International journal of human-computer studies*, 43(5–6), pp. 625–640.

Guarino, N. (1997) 'Understanding, building and using ontologies', *International Journal of Human-Computer Studies*, 46(2–3), pp. 293–310.

Guarino, N. (1998) 'The Ontological Level', *Philosophy and the cognitive sciences*.

Guarino, N., Oberle, D. and Staab, S. (2009) 'What Is an Ontology?', in *Handbook on Ontologies*. Springer, Berlin, Heidelberg (International Handbooks on Information Systems), pp. 1–17. doi: 10.1007/978-3-540-92673-3_0.

Guarino, N. and Welty, C. A. (2009) 'An overview of OntoClean', in *Handbook on ontologies*. Springer, pp. 201–220.

HermiT Reasoner: Home (no date). Available at: http://www.hermit-reasoner.com/ (Accessed: 11 July 2016).

HermiT Reasoner: Publications (no date). Available at: http://www.hermit-reasoner.com/publications.html (Accessed: 28 January 2018).

Hey, J. (2004) 'The data, information, knowledge, wisdom chain: the metaphorical link', *Intergovernmental Oceanographic Commission*, 26.

Hofweber, T. (2014) 'Logic and Ontology', in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. Fall 2014. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/fall2014/entries/logic-ontology/ (Accessed: 25 July 2017).

Horridge, M. *et al.* (2011) 'A Practical Guide To Building *OWL* Ontologies Using The Protégé-*OWL* Plugin and CO-ODE Tools Edition 1.3', *University of Manchester*.

INFÉRENCE : Définition de INFÉRENCE (no date). Available at: http://www.cnrtl.fr/definition/inf%C3%A9rence (Accessed: 27 January 2018).

ISO 15926-2:2003 - Industrial automation systems and integration -- Integration of life-cycle data for process plants including oil and gas production facilities -- Part 2: Data model (no date). Available at: https://www.iso.org/standard/29557.html (Accessed: 25 January 2018).

Jifa, G. and Lingling, Z. (2014) 'Data, DIKW, Big Data and Data Science', *Procedia Computer Science*. (2nd International Conference on Information Technology and Quantitative Management, ITQM 2014), 31, pp. 814–821. doi: 10.1016/j.procs.2014.05.332.

Jupp, S. *et al.* (2011) 'Populous: A Tool for Populating an Ontology.', in *ICBO*.

Jupp, S. *et al.* (2012) 'Populous: a tool for building *OWL* ontologies from templates', *BMC Bioinformatics*, 13(Suppl 1), p. S5. doi: 10.1186/1471-2105-13-S1-S5.

Kaneiwa, K., Iwazume, M. and Fukuda, K. (2007) 'An upper ontology for event classifications and relations', in *AI 2007: Advances in Artificial Intelligence*. Springer, pp. 394–403. Available at: http://link.springer.com/chapter/10.1007/978-3-540-76928-6_41 (Accessed: 9 February 2016).

Khatri, P. *et al.* (2004) 'Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments', *Nucleic Acids Research*, 32(Web Server issue), pp. W449-456. doi: 10.1093/nar/gkh409.

Khatri, P. *et al.* (2007) 'Onto-Tools: new additions and improvements in 2006', *Nucleic Acids Research*, 35(Web Server issue), pp. W206–W211. doi: 10.1093/nar/gkm327.

Laboratory for Applied Ontology (no date). Available at: http://www.loa.istc.cnr.it/ (Accessed: 8 August 2016).

Laboratory for Applied Ontology - DOLCE (no date). Available at: http://www.loa.istc.cnr.it/old/DOLCE.html (Accessed: 26 January 2018).

Le W3C publie les recommandations RDF et OWL (no date). Available at: https://www.w3.org/2004/01/sws-pressrelease.html.fr (Accessed: 14 January 2018).

Leal, D. (2005) 'ISO 15926', *Oil & gas science and technology*, 60(4), pp. 629–637.

'List of Reasoners | *OWL* research at the University of Manchester' (no date). Available at: http://*owl*.cs.manchester.ac.uk/tools/list-of-reasoners/ (Accessed: 28 January 2018).

McCarthy, J. (1980) 'Circumscription—a form of non-monotonic reasoning', *Artificial intelligence*, 13(1), pp. 27–39.

Musen, M. A. (2015) 'The Protégé Project: A Look Back and a Look Forward', *AI matters*, 1(4), pp. 4–12. doi: 10.1145/2757001.2757003.

Neches, R. *et al.* (1991) 'Enabling technology for knowledge sharing', *AI magazine*, 12(3), p. 36.

Noy, N. F. and McGuinness, D. L. (2000) 'Développement d'une ontologie 101: Guide pour la création de votre première ontologie', *Université de Stanford, Stanford, Traduit de l'anglais par Anila Angjeli. ht tp://www. bnf. fr/pages/infopro/normes/pdf/no-DevOnto. pdf*.

Nürnberger, A. and Wenzel, C. (2011) 'Wisdom-the blurry top of human cognition in the DIKW-model?', in *Proceedings of the EUSFLAT conference, Aix-Les-Bains, France*, pp. 584–591.

O'Connor, M. and Das, A. (2009) 'SQWRL: A Query Language for *OWL*', in *Proceedings of the 6th International Conference on OWL: Experiences and Directions - Volume 529*. Aachen, Germany, Germany: CEUR-WS.org (OWLED'09), pp. 208–215. Available at: http://dl.acm.org/citation.cfm?id=2890046.2890072 (Accessed: 26 July 2018).

'Oil Spill Ontologies | ENVISION - ENVIronmental Services Infrastructure with ONtologies' (no date). Available at: http://www.envision-project.eu/resources/ontologies/oil-spill-ontologies/ (Accessed: 15 February 2016).

ONTOLOGIE : Définition de ONTOLOGIE (no date). Available at: http://www.cnrtl.fr/definition/ontologie (Accessed: 25 July 2017).

Ontology (Computer Science) - definition in Encyclopedia of Database Systems (no date). Available at: http://tomgruber.org/writing/ontology-definition-2007.htm (Accessed: 22 January 2018).

OWL - Semantic Web Standards (no date). Available at: https://www.w3.org/*OWL*/ (Accessed: 21 January 2018).

OWL 2 Web Ontology Language Document Overview (Second Edition) (2012). Available at: https://www.w3.org/TR/owl2-overview/ (Accessed: 23 November 2016).

OWL 2 Web Ontology Language Primer (Second Edition) (no date). Available at: https://www.w3.org/TR/owl2-primer/#Object_Properties (Accessed: 9 July 2016).

OWL 2 Web Ontology Language Profiles (Second Edition) (no date). Available at: https://www.w3.org/TR/owl2-profiles/#*OWL*_2_RL (Accessed: 4 August 2018).

OWL Web Ontology Language Overview (no date). Available at: https://www.w3.org/TR/2004/REC-*owl*-features-20040210/#s1.2 (Accessed: 21 January 2018).

Pascal Gribomont (no date). Available at: http://www.montefiore.ulg.ac.be/~gribomon/ (Accessed: 25 January 2018).

Patton, M. Q. (1999) 'Enhancing the quality and credibility of qualitative analysis', *Health Services Research*, 34(5 Pt 2), pp. 1189–1208.

Poli, R. (1999) 'Framing ontology', *Available online from the Ontology resource guide for philosphers: http://www. formalontology. it/essays/Framing. pdf*, 23, pp. 19–26.

Provitolo, D., Dubos-Paillard, E. and Müller, J.-P. (2009) 'Vers une ontologie des risques et des catastrophes: le modèle conceptuel', in *Ontologie et dynamique des systèmes complexes, perspectives interdisciplinaires*, pp. 1–16. Available at: https://halshs.archives-ouvertes.fr/halshs-00643597/ (Accessed: 20 June 2016).

Psarros, G., Skjong, R. and Eide, M. S. (2010) 'Under-reporting of maritime accidents', *Accident Analysis & Prevention*, 42(2), pp. 619–625. doi: 10.1016/j.aap.2009.10.008.

Raimbault, T. *et al.* (2008) 'Une synthèse des modèles de représentation des connaissances à base de Graphes Conceptuels et *OWL*', *Revue des nouvelles technologies de l'information*, (12), pp. 45–66.

RDF 1.1 Concepts and Abstract Syntax (no date). Available at: https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/#section-rdf-graph (Accessed: 21 January 2018).

RDF 1.1 Semantics (no date). Available at: https://www.w3.org/TR/rdf11-mt/ (Accessed: 22 January 2018).

RDF Primer (no date). Available at: https://www.w3.org/TR/2004/REC-rdf-primer-20040210/#ref-rdf-concepts (Accessed: 14 January 2018).

RDF Schema 1.1 (no date). Available at: https://www.w3.org/TR/2014/REC-rdf-schema-20140225/ (Accessed: 14 January 2018).

RDF Semantics (no date). Available at: https://www.w3.org/TR/2004/REC-rdf-mt-20040210/ (Accessed: 14 January 2018).

Resource Description Framework (RDF): Concepts and Abstract Syntax (no date). Available at: https://www.w3.org/TR/rdf-concepts/#section-data-model (Accessed: 14 January 2018).

Robert, F. (2006) 'Science et ontologie', *Archives de Philosophie*, 69(1), pp. 101–122.

Roche, C. (2005) 'Terminologie et ontologie, Abstract', *Langages*, (157), pp. 48–62. doi: 10.3917/lang.157.0048.

Roussey, C. *et al.* (2011) 'An Introduction to Ontologies and Ontology Engineering', in Falquet, G. et al., *Ontologies in Urban Development Projects*. London: Springer London, pp. 9–38. doi: 10.1007/978-0-85729-724-2_2.

Rowley, J. (2007) 'The wisdom hierarchy: representations of the DIKW hierarchy', *Journal of Information Science*, 33(2), pp. 163–180. doi: 10.1177/0165551506070706.

Rychlỳ, P. (2008) 'A lexicographer-friendly association score', *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, 2008, p. 6.

Sánchez, D. M., Cavero, J. M. and Martínez, E. M. (2007) 'The road toward ontologies', in *Ontologies*. Springer, pp. 3–20.

Scherp, A. *et al.* (2009) 'F–a model of events based on the foundational ontology dolce+ *DnS UL*tralight', in *Proceedings of the fifth international conference on*

Knowledge capture. ACM, pp. 137–144. Available at: http://dl.acm.org/citation.cfm?id=1597760 (Accessed: 15 August 2017).

ScienceDirect.com | Science, health and medical journals, full text articles and books. (no date). Available at: http://www.sciencedirect.com/ (Accessed: 28 September 2017).

SÉMIOTIQUE : Définition de SÉMIOTIQUE (no date). Available at: http://www.cnrtl.fr/lexicographie/s%C3%A9miotique (Accessed: 4 January 2018).

Shaw, R., Troncy, R. and Hardman, L. (2009) 'LODE: Linking Open Descriptions of Events', *School of Information*. Available at: http://escholarship.org/uc/item/4pd6b5mh (Accessed: 8 August 2016).

Sirin, E. *et al.* (2007) 'Pellet: A practical *owl*-dl reasoner', *Web Semantics: science, services and agents on the World Wide Web*, 5(2), pp. 51–53.

Sitzia, J. (1999) 'How valid and reliable are patient satisfaction data? An analysis of 195 studies', *International Journal for Quality in Health Care*, 11(4), pp. 319–328. doi: 10.1093/intqhc/11.4.319.

Smith, B. (2006) 'Against idiosyncrasy in ontology development', *Frontiers in Artificial Intelligence and Applications*, 150, p. 15.

SPARQL 1.1 Overview (no date). Available at: https://www.w3.org/TR/sparql11-overview/ (Accessed: 28 January 2018).

Stoan, S. K. (1984) 'Research and library skills: An analysis and interpretation', *College & research libraries*, 45(2), pp. 99–109.

Studer, R., Benjamins, V. R. and Fensel, D. (1998) 'Knowledge engineering: Principles and methods', *Data & Knowledge Engineering*, 25(1), pp. 161–197. doi: 10.1016/S0169-023X(97)00056-6.

SUBSOMPTION : Définition de SUBSOMPTION (no date). Available at: http://www.cnrtl.fr/definition/subsomption (Accessed: 4 January 2018).

Suchomel, V. and Pomikálek, J. (2012) 'Efficient web crawling for large text corpora', in *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pp. 39–43.

SWRL: A Semantic Web Rule Language Combining OWL and RuleML (no date). Available at: https://www.w3.org/Submission/SWRL/ (Accessed: 8 July 2018).

swrlapi: Java API for working with the SWRL rule and SQWRL query languages (2018). Protege Project. Available at: https://github.com/protegeproject/swrlapi (Accessed: 8 July 2018).

Tableau Public (no date) *Tableau Public*. Available at: https://public.tableau.com/en-us/s/ (Accessed: 27 July 2018).

TAXINOMIE : Définition de TAXINOMIE (no date). Available at: http://www.cnrtl.fr/definition/taxinomie (Accessed: 4 January 2018).

The OBO Foundry (no date). Available at: http://www.obofoundry.org/ (Accessed: 28 January 2018).

'Tom Gruber | AI product designer' (no date). Available at: http://tomgruber.org/ (Accessed: 29 September 2018).

Venkataraman, P. and Mendonca, D. (2003) 'Using process data to populate ontologies', in *IEEE International Conference on Systems, Man and Cybernetics, 2003. IEEE International Conference on Systems, Man and Cybernetics, 2003*, pp. 2156–2161 vol.3. doi: 10.1109/ICSMC.2003.1244203.

Vessel details for: MAERSK GARONNE (Container Ship) - IMO 9235579, MMSI 219219000, Call Sign OUTK2 Registered in Denmark | AIS Marine Traffic (no date) *MarineTraffic.com*. Available at: http://www.marinetraffic.com/en/ais/details/ships/shipid:156491/mmsi:219219000/imo:9235579/vessel:MAERSK_GARONNE (Accessed: 28 January 2018). e-CFR

Vigneron, J., Rallo, J.-M. and Guarnieri, F. (2014) *Apports des ontologies à la création de bases de connaissances pour la maîtrise des conformités légales en santé et sécurité au travail*. report. MINES ParisTech, p. 8 p. Available at: https://hal-mines-paristech.archives-ouvertes.fr/hal-00990835/document (Accessed: 15 August 2017).

What is Cytoscape? (no date). Available at: http://www.cytoscape.org/what_is_cytoscape.html (Accessed: 28 December 2017).

Wolstencroft, K. *et al.* (2011) 'RightField: embedding ontology annotation in spreadsheets', *Bioinformatics*, 27(14), pp. 2021–2022. doi: 10.1093/bioinformatics/btr312.

Wolstencroft, K. *et al.* (2012) 'RightField: Scientific Knowledge Acquisition by Stealth through Ontology-Enabled Spreadsheets', in *Knowledge Engineering and Knowledge Management. International Conference on Knowledge Engineering and Knowledge Management*, Springer, Berlin, Heidelberg (Lecture Notes in Computer Science), pp. 438–441. doi: 10.1007/978-3-642-33876-2_42.

Wolstencroft, K., Owen, S., Krebs, O., *et al.* (2013) 'Semantic Data and Models Sharing in Systems Biology: The Just Enough Results Model and the SEEK Platform', in *The Semantic Web – ISWC 2013. International Semantic Web Conference*, Springer, Berlin, Heidelberg (Lecture Notes in Computer Science), pp. 212–227. doi: 10.1007/978-3-642-41338-4_14.

Wolstencroft, K., Owen, S., Horridge, M., *et al.* (2013) 'Stealthy annotation of experimental biology by spreadsheets', *Concurrency and Computation: Practice and Experience*, 25(4), pp. 467–480. doi: 10.1002/cpe.2941.

Admiral Thad W. Allen > U.S. Department Of Defense > Biography View (no date). Available at: https://www.defense.gov/About/Biographies/Biography-View/Article/602793/ (Accessed: 22 June 2018).

Allen, Thad W. (2010) *National Incident Commander's Report: MC252 Deepwater Horizon*, p. 28.

blowout - Schlumberger Oilfield Glossary (no date). Available at: https://www.glossary.oilfield.slb.com/en/Terms/b/blowout.aspx (Accessed: 20 October 2018).

Bly, M. and BP (2010) *Deepwater Horizon accident investigation report*. Diane Publishing, p. 369. Available at: http://books.google.com/books?hl=en&lr=&id=oJnW9R4m_3sC&oi=fnd&pg=PA9&dq=%22work+was%22+%22people+lost+their+lives,+and+17+others+were+injured.+The+fire,+which+was+fed+by%22+%22from+other+BP+spill+response+activities+and+organizations.+The+ability+to%22+&ots=nXpL_fDUwW&sig=zlbmirPSd8BTg8k4ap3HpxduCow (Accessed: 13 February 2015).

'BP EP - 2009 - Initial Exploration Plan, Mississippi Canyon Block' (2009).

Camilli, R. *et al.* (2012) 'Acoustic measurement of the Deepwater Horizon Macondo well flow rate', *Proceedings of the National Academy of Sciences of the United States of America*, 109(50), p. 6. doi: 10.1073/pnas.1100385108.

Dahlstrom, D. J. and Carter, J. T. V. (2013) 'Inverse Modeling with PEST++ and GENIE', *Ground Water*, p. n/a-n/a. doi: 10.1111/gwat.12021.

Deepwater Horizon Study Group (2011) *Final Report on the Investigation of the Macondo Well Blowout Deepwater Horizon Study Group*, p. 126.

Det Norske Veritas (2011) *Final Report For United States Department Of The Interior Bureau Of Ocean Energy Management, Regulation, And Enforcement, Forensic Examination Of Deepwater Horizon Blowout Preventer*, P. 581.

Det Norske Veritas Final Report For United States Department Of The Interior Bureau Of Ocean Energy Management, Regulation, And Enforcement Washington, Dc 20240 Forensic Examination Of Deepwater Horizon Blowout Preventer Volume Ii Appendices (2011).

Eude, T., Napoli, A. and Guarnieri, F. (2016) 'A thorough analysis of the engineering solutions deployed to stop the oil spill following the deepwater horizon disaster', *Chemical Engineering Transactions*, p. 6. doi: 10.3303/CET1648130.

Greenemeier, L. (2010) *Drill BP, Drill: By Boring Relief Wells Closer to the Oil Reservoir BP Hopes to Up Odds of Success*, *Scientific American*. Available at: https://www.scientificamerican.com/article/bp-relief-well-drilling/ (Accessed: 19 August 2017).

Hickman, S. H. *et al.* (2012) 'Scientific basis for safely shutting in the Macondo Well after the April 20, 2010 Deepwater Horizon blowout', *Proceedings of the National Academy of Sciences*, 109(50), p. 6. doi: 10.1073/pnas.1115847109.

Hsieh, P. (2010) *Computer simulation of reservoir depletion and oil flow from the Macondo well following the Deepwater Horizon blowout*, p. 22. Available at:

http://home.doi.gov/deepwaterhorizon/upload/FRTG-report-Appendix-A-Hsieh-Reservoir-Simulation-Report-updated.pdf (Accessed: 2 February 2016).

Judge Barbier (2015) *Findings Of Fact And Conclusions Of Law Phase Two Trial*.

Loss of Well Control | Bureau of Safety and Environmental Enforcement (no date). Available at: https://www.bsee.gov/stats-facts/offshore-incident-statistics/loss-of-well-control (Accessed: 13 July 2017).

Lubchenco, J. *et al.* (2012) 'Science in support of the Deepwater Horizon response', *Proceedings of the National Academy of Sciences*, 109(50), p. 10. doi: 10.1073/pnas.1204729109.

McNutt, M. K., Chu, S., *et al.* (2012) 'Applications of science and engineering to quantify and control the Deepwater Horizon oil spill', *Proceedings of the National Academy of Sciences*, 109(50), p. 8. doi: 10.1073/pnas.1214389109.

McNutt, M. K., Camilli, R., *et al.* (2012) 'Review of flow rate estimates of the Deepwater Horizon oil spill', *Proceedings of the National Academy of Sciences*, 109(50), p. 8. doi: 10.1073/pnas.1112139108.

MDL 2179 Trial Docs - Phase One (no date). Available at: http://www.mdl2179trialdocs.com/index.php?page=phase1 (Accessed: 31 July 2018).

MDL 2179 Trial Docs - Phase Three (no date). Available at: http://www.mdl2179trialdocs.com/index.php?page=phase3 (Accessed: 31 July 2018).

MDL 2179 Trial Docs - Phase Two (no date). Available at: http://www.mdl2179trialdocs.com/index.php?page=phase2 (Accessed: 31 July 2018).

MediaWiki/fr - MediaWiki (no date). Available at: https://www.mediawiki.org/wiki/MediaWiki/fr (Accessed: 29 January 2018).

Methane Hydrate | Department of Energy (no date). Available at: https://www.energy.gov/fe/science-innovation/oil-gas-research/methane-hydrate (Accessed: 25 March 2018).

National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling (2011a) *Deep water: the Gulf oil disaster and the future of offshore drilling : report to the President*. [Washington, D.C.]: National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling : For sale by the Supt. of Docs., U.S. G.P.O., p. 398.

National Commission On The BP Deepwater Horizon Oil Spill And Offshore Drilling (2011b) *Stopping The Spill: The Five-Month Effort To Kill The Macondo Well ---Draft--- Staff Working Paper No. 6*, P. 39.

National Incident Command, I. S. G., Flow Rate Technical Group (2011) 'Assessment of Flow Rate Estimates for the Deepwater Horizon / Macondo Well Oil Spill'.

'PEST Model-Independent Parameter Estimation User Manual, 5th Edition.' (no date), p. 336.

PURGATOIRE : Définition de PURGATOIRE (no date). Available at: http://www.cnrtl.fr/definition/purgatoire (Accessed: 18 August 2018).

Risk Assessment Of The Deepwater Horizon Blowout Preventer (Bop) Control System April 2000 - Final Report (2000).

SINTEF Offshore Blowout Database (no date) *SINTEF*. Available at: https://www.sintef.no/en/projects/sintef-offshore-blowout-database/ (Accessed: 13 July 2017).

The Federal Interagency Solutions Group, *et al.* (2010) *Deepwater Horizon oil budget calculator*. Available at: http://www.eoearth.org/view/article/161891/ (Accessed: 13 February 2015).

The Federal Interagency Solutions Group (2010) *Deepwater Horizon oil budget calculator*, p. 217. Available at: http://www.eoearth.org/view/article/161891/ (Accessed: 13 February 2015).

United States Coast Guard (2011) *Federal On Scene Coordinator Report to the National Response Team*, p. 222.

US Chemical Safety And Hazard Investigation Board (2014) *Investigation Report Explosion And Fire At The Macondo Well*, P. 162.

USGS Scientist Honored with Prestigious Federal Employee of the Year Medal for Role in Ending Deepwater Horizon Oil Spill (no date). Available at: https://soundwaves.usgs.gov/2011/11/awards.html (Accessed: 2 September 2018).

USGS.gov | Science for a changing world (no date). Available at: https://www.usgs.gov/ (Accessed: 1 September 2018).

WashingtonsBlog (no date) 'BP Oil Spill: Case NOT Closed | Washington's Blog'. Available at: http://washingtonsblog.com/2012/10/bp-oil-spill-case-not-closed.html (Accessed: 16 March 2018).

well control - Schlumberger Oilfield Glossary (no date). Available at: https://www.glossary.oilfield.slb.com/en/Terms/w/well_control.aspx (Accessed: 4 November 2018).

About WordNet - WordNet - About WordNet (no date). Available at: https://wordnet.princeton.edu/ (Accessed: 26 January 2018).

Accueil | Corpus arboré pour le français (no date). Available at: http://ftb.linguist.univ-paris-diderot.fr/ (Accessed: 10 February 2018).

'Adam Kilgarriff bibliography | Sketch Engine' (no date). Available at: https://www.sketchengine.co.uk/adam-kilgarriff-structured-bibliography/ (Accessed: 18 November 2017).

Adposition (2015) *SIL Glossary of Linguistic Terms*. Available at: https://glossary.sil.org/term/adposition (Accessed: 14 February 2018).

Barbet, C. and Saussure, L. de (2012) 'Présentation : Modalité et évidentialité en français', *Langue française*, (173), pp. 3–12. doi: 10.3917/lf.173.0003.

Cailliau, F. (2010) *Des ressources aux traitements linguistiques: le rôle d 'une architecture linguistique*. PhD Thesis. Université Paris-Nord-Paris XIII.

Cohen, K. B. *et al.* (2017) 'Coreference annotation and resolution in the Colorado Richly Annotated Full Text (CRAFT) corpus of biomedical journal articles', *BMC Bioinformatics*, 18. doi: 10.1186/s12859-017-1775-9.

'Concordance | Sketch Engine' (no date). Available at: https://www.sketchengine.co.uk/user-guide/user-manual/concordance-introduction/ (Accessed: 18 November 2017).

Corpus info: French Web 2012 (frTenTen12) (no date). Available at: https://the.sketchengine.co.uk/corpus/corp_info?corpname=preloaded/frtenten12_1&struct_attr_stats=1&subcorpora=1 (Accessed: 18 November 2017).

van Deemter, K. and Kibble, R. (1999) 'What is coreference, and what should coreference annotation be?', in. Association for Computational Linguistics, p. 90. doi: 10.3115/1608810.1608828.

DÉNOTATION : Définition de DÉNOTATION (no date). Available at: http://www.cnrtl.fr/definition/d%C3%A9notation (Accessed: 18 March 2018).

Djemaa, M. *et al.* (2016) 'Corpus annotation within the French FrameNet: a domain-by-domain methodology', in *Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia. Available at: https://hal.archives-ouvertes.fr/hal-01391526 (Accessed: 18 March 2018).

Dunietz, J., Levin, L. and Carbonell, J. (2015) 'Annotating Causal Language Using Corpus Lexicography of Constructions', in. Association for Computational Linguistics, pp. 188–196. doi: 10.3115/v1/W15-1622.

Dunietz, J., Levin, L. and Carbonell, J. (2017a) 'Automatically Tagging Constructions of Causation and Their Slot-Fillers', *Transactions of the Association for Computational Linguistics*, 5, pp. 117–133.

Dunietz, J., Levin, L. and Carbonell, J. (2017b) 'The BECauSE Corpus 2.0: Annotating Causality and Overlapping Relations', in *Proceedings of the 11th Linguistic Annotation Workshop*, pp. 95–104.

Eckart de Castilho, R. *et al.* (2014) 'WebAnno: a flexible, web-based annotation tool for CLARIN', *Proceedings of the CLARIN Annual Conference (CAC) 2014*.

‘French Web corpus (frTenTen) search | Sketch Engine’ (no date). Available at: https://www.sketchengine.co.uk/frtenten-corpus/ (Accessed: 18 November 2017).

HYPONYMIE : Définition de HYPONYMIE (no date). Available at: http://www.cnrtl.fr/definition/hyponymie (Accessed: 7 March 2018).

Jakubíček, M. *et al.* (2013) ‘The tenten corpus family’, in *7th International Corpus Linguistics Conference CL*, pp. 125–127.

Kilgarriff, A. *et al.* (2015) ‘Longest–commonest Match’, in *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*, pp. 11–13.

L’analyseur syntaxique FRMG pour le français | FRMG Wiki (no date). Available at: http://alpage.inria.fr/frmgwiki/wiki/lanalyseur-syntaxique-frmg-pour-le-fran%C3%A7ais (Accessed: 10 February 2018).

Lee, H. *et al.* (2013) ‘Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules’, *Computational Linguistics*, 39(4), pp. 885–916. doi: 10.1162/COLI_a_00152.

LU index | fndrupal (no date). Available at: https://framenet.icsi.berkeley.edu/fndrupal/luIndex (Accessed: 24 February 2018).

Luo, Z. *et al.* (2016) ‘Commonsense Causal Reasoning between Short Texts’, p. 10.

Manguin, J.-L. *et al.* (2004) ‘Le dictionnaire électronique des synonymes du CRISCO. Un mode d’emploi à trois niveaux’. Available at: http://www.crisco.unicaen.fr/Le-dictionnaire-electronique-des-synonymes.html (Accessed: 1 April 2017).

Manguin, J.-L. (2004) ‘Transitivité partielle de la synonymie : application aux dictionnaires de synonymes’, *Corela. Cognition, représentation, langage*, (2–2). doi: 10.4000/corela.611.

Mirza, P. *et al.* (2014) ‘Annotating causality in the TempEval-3 corpus’, in *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pp. 10–19.

Morel, M. and François, J. (2015) ‘Le Dictionnaire Electronique des Synonymes du CRISCO : un outil de plus en plus interactif’, *Revue française de linguistique appliquée*, XX(1), pp. 9–28.

nsubj (no date). Available at: http://universaldependencies.org/fr/dep/nsubj.html (Accessed: 15 May 2018).

Passage: ANR MDCA Passage (no date). Available at: http://atoll.inria.fr/passage/ (Accessed: 12 February 2018).

Penn Treebank II Tags (no date) *Gist*. Available at: https://gist.github.com/nlothian/9240750 (Accessed: 10 February 2018).

Penn Treebank P.O.S. Tags (no date). Available at: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html (Accessed: 1 February 2018).

Peters, M. E. *et al.* (2018) ‘Deep contextualized word representations’, *arXiv:1802.05365 [cs]*. Available at: http://arxiv.org/abs/1802.05365 (Accessed: 13 May 2018).

PHRASE : Définition de PHRASE (no date). Available at: http://www.cnrtl.fr/definition/phrase (Accessed: 16 February 2018).

Ploux, S. and Victorri, B. (1998a) 'Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes', *Traitement automatique des langues*, (39), pp. 161–182.

Ploux, S. and Victorri, B. (1998b) 'Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes', *Traitement automatique des langues*, (39), pp. 161–182.

Popescu-Belis, A. (1998) 'How corpora with annotated coreference links improve reference resolution', *First Int. Conf. on Language Resources and Evaluation*, pp. 567–572.

Popescu-Belis, A., Robba, I. and Sabah, G. (1998) 'Reference Resolution beyond Coreference: a Conceptual Frame and its Application', *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*. Available at: http://www.aclweb.org/anthology/P98-2172.

Présentation de Semantic MediaWiki — semantic-mediawiki.org (no date) *Semantic MediaWiki*. Available at: https://www.semantic-mediawiki.org/wiki/Help:Introduction_to_Semantic_MediaWiki/fr (Accessed: 29 January 2018).

Purnelle, G. (1996) 'Utilisation d'une banque de données de textes latins lemmatisés et analysés. Problèmes spécifiques aux données linguistiques', in *Bases de données linguistiques: conceptions, réalisations, exploitations. Actes du Colloque International de Corte (11–14 octobre 1995)*, pp. 295–307.

Reboul, A. and Gaiffe, B. (1999) 'Représentations mentales et référence'. Available at: https://halshs.archives-ouvertes.fr/halshs-00003843/document (Accessed: 22 March 2018).

Santorini, B. (1990) 'Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision)', *Technical Reports (CIS)*, p. 570.

Schütze, H. and Pedersen, J. O. (1995) *Information Retrieval Based on Word Senses*.

'Sketch Engine | language corpus management and query system' (no date). Available at: https://www.sketchengine.co.uk/ (Accessed: 18 November 2017).

'Statistics used in Sketch Engine | Sketch Engine' (no date). Available at: https://www.sketchengine.co.uk/documentation/statistics-used-in-sketch-engine/ (Accessed: 18 November 2017).

SYNTAGME : Définition de SYNTAGME (no date). Available at: http://www.cnrtl.fr/definition/syntagme (Accessed: 26 December 2017).

The Proposition Bank (PropBank) (no date). Available at: https://propbank.github.io/ (Accessed: 24 February 2018).

Toutanova, K. *et al.* (2003) 'Feature-rich part-of-speech tagging with a cyclic dependency network', in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, pp. 173–180.

Toutanova, K. and Manning, C. D. (2000) 'Enriching the knowledge sources used in a maximum entropy part-of-speech tagger', in *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large*

corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13. Association for Computational Linguistics, pp. 63–70.

TreeTagger (no date). Available at: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (Accessed: 16 January 2018).

TreeTagger - a language independent part-of-speech tagger | Institute for Natural Language Processing | University of Stuttgart (no date). Available at: http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.en.html (Accessed: 16 January 2018).

UD_English (no date). Available at: http://universaldependencies.org/treebanks/en/index.html (Accessed: 9 February 2018).

Unified Verb Index (no date). Available at: https://verbs.colorado.edu/verb-index/vn3.3/ (Accessed: 18 March 2018).

UzZaman, N. *et al.* (2013) 'SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations', in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA: Association for Computational Linguistics, pp. 1–9. Available at: http://www.aclweb.org/anthology/S13-2001 (Accessed: 31 May 2018).

Verb Sense Annotation Project (no date). Available at: http://verbs.colorado.edu/VSAP/annot.html (Accessed: 21 March 2018).

Victorri, B. (2002) 'Espaces_semantiques_et_representation_du_sens.doc'.

Victorri, B. and Fuchs, C. (1996a) *La polysémie-Construction dynamique du sens*. Hermes. Available at: https://halshs.archives-ouvertes.fr/halshs-00713735/ (Accessed: 6 April 2017).

Victorri, B. and Fuchs, C. (1996b) *La polysémie-Construction dynamique du sens*. Hermes. Available at: https://halshs.archives-ouvertes.fr/halshs-00713735/ (Accessed: 3 April 2017).

Welcome to FrameNet! | fndrupal (no date). Available at: https://framenet.icsi.berkeley.edu/fndrupal/ (Accessed: 21 March 2018).

With - English Grammar Today - Cambridge Dictionary (no date). Available at: https://dictionary.cambridge.org/grammar/british-grammar/with (Accessed: 14 May 2018).

wngloss(7WN) | WordNet (no date). Available at: https://wordnet.princeton.edu/documentation/wngloss7wn (Accessed: 7 March 2018).

'Word Sketch Difference - compare collocations | Sketch Engine' (no date). Available at: https://www.sketchengine.co.uk/user-guide/user-manual/word-sketch-difference/ (Accessed: 23 November 2017).

WordNet (no date) *MIT Press*. Available at: https://mitpress.mit.edu/books/wordnet (Accessed: 20 February 2018).

WordNet Search - 3.1 (no date). Available at: http://wordnetweb.princeton.edu/perl/webwn?o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&r=1&s=accident&i=35&h=1113000221302213013022221302213020110000000#c (Accessed: 26 January 2018).

Glossaire

3-ram capping stack : le *3-ram capping stack* (*CS*) ou encore *sealing cap* est un équipement technqiue *ad-hoc* déployé au-dessus du *LMRP* et qui permettra la fermeture du puits Macondo le 15 juillet 2010.

Barrel Per Day (***BPD***) : unité de mesure du débit de fuite. 1 *BPD* correspond à environ 159 litres/jour.

BlowOut Preventer ou **Bloc Obturateur de Puits** (**BOP**) : équipement technique pour les opérations de *well control*, utilisé à la fois en opérations de routine et en barrière de sécurité ; il est installé sur la tête de puits (Deepwater Horizon Study Group, 2011, pp. 25–28). Une description précise de cet organe critique est donnée dans le rapport de BP (Appendix H: description of the BOP Stack and Control System Bly and BP, 2010) et du DNV (Det Norske Veritas, 2011, pp. 14–17)

Blowout : dans les opérations de forage, « écoulement incontrôlé des fluides de formation [géologique] provenant d'un puits. » (*blowout - Schlumberger Oilfield Glossary*, no date).

Classifieur : dans le domaine de l'apprentissage automatique, algorithme capable de prédire la classe à laquelle appartient un objet en fonction des caractéristiques de ce dernier. On parle de classification quand les données sont discrètes et de régression quand les données sont continues. Type de modèle de *machine learning* permettant de différencier deux classes discrètes ou plus. (*Glossaire du machine learning*, no date).

Complétion : opération de forage où il s'agit de sceller le puits avec du ciment pour le mettre en sécurité, puis de le tester en pression. Cette opération permet à la plateforme de forage de se déconnecter et de laisser la place par la suite à une plateforme de production, qui une fois connectée, « brisera le scellé » et commencera l'extraction.

Dénotation : la dénotation est ce qui correspond à l'extension ou l'ensemble de des sèmes génériques d'une unité lexicale. C'est l'ensemble des traits distinctifs qui objectivement caractérisent cette classe. ("DÉNOTATION : Définition de DÉNOTATION," n.d.).

Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) : *DOLCE* est une ontologie de haut niveau créée par Gangemi (2002).

Disque de rupture : équipements de sécurité installés à intervalles réguliers en profondeur et dont le but est justement de se rompre à une pression donnée pour protéger le puits d'une surpression interne.

Federal On-Scene Coordinator (FOSC) : chef des opérations sur site, par délégation de pouvoir du *NIC*.

Flow Rate Technical Group : groupe *ad-hoc* de scientifiques du gouvernement américain en charge d'apporter une estimation du débit de fuite du puits Macondo.

Government-Led Science Team (***GLST***) : l'équipe de scientifiques de très haut niveau, dirigée par le Pr. Steven Chu, chargée de réfléchir aux solutions à mettre en œuvre pour reprendre le contrôle sur le puits.

Intégrité : capacité physique (mécanique) d'un puits à supporter l'éruption, à contenir l'effluent, et à ne pas s'effondrer sur lui-même.
Intelligence Artificielle (IA) : l'intelligence artificielle est la capacité d'une machine à reproduire le comportement ou l'intellection d'un humain dans un environnement donné.
Junk Shot Manifold : c'est une « clarinette », un équipement de *piping* qui permet de distribuer des flux entre différents équipements, l'analogie peut être faite avec un boîtier de distribution électrique. C'est un équipement nécessaire pour la mise en œuvre d'une procédure spécifique (*top kill*) pour tenter de tuer le puits.
Lemme : un lemme est une « forme graphique choisie conventionnellement comme adresse dans un lexique » (Ploux and Victorri, 1998a).
Lower Marine Riser Pack (***LMRP***) : le *LMRP* constitue la partie supérieure du *BOP* et assure à la fois la connexion entre le riser et le *BOP* et aussi une fonction de *well control* par le biais d'équipements de contrôle de flux dans l'annulaire.
Machine Comprehension **(*MC*)** : les modèles de « *Machine Comprehension* » répondent aux questions en langage naturel en sélectionnant une plage de réponse dans un texte de données probantes [preuves].
Machine Learning (***ML***) : apprentissage automatique (ou machine). « Programme ou système qui crée (entraîne) un modèle prédictif à partir de données d'entrée. Le système utilise le modèle entraîné pour effectuer des prédictions utiles à partir de nouvelles données (jamais vues auparavant) issues de la même distribution que celle utilisée pour entraîner le modèle. » (*Glossaire du machine learning*, no date).
Marée noire : expression désignant un « déversement accidentel d'hydrocarbures en mer » entraînant une pollution de l'environnement.
Morphème : « Un morphème est une unité linguistique minimale ayant une forme et un sens » (Tellier, no date, p. 23)
Named Entity Recognizer (***NER***) : type d'algorithme capable de détecter des expressions écrites susceptibles de représenter des instances appartenant à des classes prédéfinies.
National Incident Commander (NIC) : échelon le plus élevé du commandement dans l'intervention. Le *NIC* est nommé par le président des Etats-Unis lorsque le « niveau d'alerte » de pollution est caractérisé en *SONS*. D'après Thad Allen, alors le premier *NIC* de l'histoire des Etats-Unis, il s'agissait avant tout de « *promouvoir la cohésion de l'effort dans l'ensemble de l'administration – « prendre mille points lumineux et les transformer en faisceau laser »* » (2010).
Natural Language Processing **(*NLP*)** : le **Traitement Automatique du Langage Naturel (TALN)** est l'ensemble des outils algorithmiques qui permettent à une machine d'appréhender le texte écrit en langage naturel pour des utilisations en Intelligence Artificielle.
Natural Language Understanding **(*NLU*)** : phase avancée du TALN qui donne à la machine la possbilité de se rapprocher du raisonnement humain face à la lecture d'un texte.
Pipe Buckling Effective Compression : phénomène de déformation plastique subie par la tige de forage lors de l'éruption et du passage du mélange boue-gaz à

grande vitesse dans le BOP, qui a empêché les mâchoires du BOP de se fermer correctement (US Chemical Safety And Hazard Investigation Board, 2014, Para. 3.2.3 The AMF/Deadman Fails To Seal The Well: Buckled Drillpipe).

Puis réactif : puits qui entre en éruption naturellement parce que le différentiel de pression entre la tête de puits et l'environnement est positif.

Purgatoire : le purgatoire est « Selon la Tradition, lieu (étant représenté souvent comme enflammé) où les baptisés, morts en état de grâce, mais non entièrement purifiés par la pénitence des traces de leurs péchés, achèvent leur purification avant la vision béatifique » Définition A donnée par le CNRTL (*PURGATOIRE : Définition de PURGATOIRE*, no date)

Riser : le *riser* est le flexible qui relie la plateforme de forage au puits et dans lequel est descendue la tige de forage. Les fluides de contrôle du puits y circulent également pendant les opérations de forage. A son extrémité arrachée et tombée au fond de l'eau sera découverte une des deux fuites d'où s'échappent les hydrocarbures du puits.

Sémantique distribuée : l'hypothèse de la sémantique distribuée consiste à poser comme vrai que « des mots avec des significations similaires apparaîtront avec des [mots] voisins similaires s'il y a suffisamment de texte disponible » (Schütze and Pedersen, 1995). Il y a des associations de mots (co-occurrences) plus fréquentes que d'autres qui composent les expressions et les phrases.

Spill Of National Significance SONS : selon le *CFR title 40 part 300*, une pollution de type *SONS* est un « *déversement qui, en raison de sa gravité, de son ampleur, de son emplacement, de son impact réel ou potentiel sur la santé publique et le bien-être ou l'environnement, ou de l'effort d'intervention nécessaire, est si complexe qu'il exige une coordination extraordinaire des ressources fédérales, étatiques, locales et des parties responsables pour contenir et nettoyer le rejet.* »

Syntagme : « Un syntagme est [...] un groupe de mots qui correspond à un sous-arbre d'un arbre d'analyse syntaxique complet. » (Tellier, no date, p. 38). Ou encore « une combinaison de morphèmes ou de mots qui se suivent et produisent un sens acceptable », particulièrement « un groupe d'unités linguistiques significatives formant une unité dans une organisation hiérarchisée de la phrase. » , définition B (*SYNTAGME : Définition de SYNTAGME*, no date)

Traitement Automatique du Langage Naturel (TALN) : voir *NLP*

Triple : le *triple* est la relation ontologique irréductible constituée de trois parties : sujet, prédicat, objet.

Tuer un puits : « tuer un puits » est une expression utilisée pour décrire l'opération qui permet de reprendre le contrôle d'un puits rentré en éruption d'une part par l'injection de grands volumes de boue pour rééquilibrer la pression dans le réservoir géologique puis d'autre part par l'injection de ciment pour sceller définitivement le puits.

Unified Command : la structure de commandement de l'intervention. Pour bien comprendre les rôles et les responsabilités au sein de *l'Unified Command*, consulter les rapports (Allen, Thad W., 2010; United States Coast Guard, 2011).

Well control : les opérations de contrôle du puits ont pour objectif de « maintenir la pression sur les formations [géologiques] ouvertes (c'est-à-dire exposées au puits de

forage) pour empêcher ou diriger l'écoulement des fluides de formation dans le puits de forage » (*well control - Schlumberger Oilfield Glossary*, no date).

Table des figures

Table des tableaux

Table des équations

Table des annexes

Annexes

Annexe 1 : analyse des « *patterns* » du lemme « accident » dans le corpus frTenTen12

modifier	**90785**	**15,34**	**et_ou**	**108287**	**18,3**
----------			----------		
vasculaire	13986	11,69	incident	3469	8,45
mortel	9294	10,46	maladie	11868	8,11
grave	6345	8,75	infarctus	501	6,98
nucléaire	5040	8,64	suicide	894	6,72
domestique	3152	8,63	panne	780	6,6
ischémique	927	8,31	décès	999	6,5
corporel	2618	8,3	maternité	377	6,06
tragique	977	7,4	chute	736	5,78
cérébrovasculaires	323	6,85	embouteillage	227	5,78
ferroviaire	711	6,84	invalidité	253	5,74
majeur	2272	6,79	hypertension	237	5,43
automobile	1003	6,66	attentat	328	5,37
coronarien	282	6,49	pollution	404	5,33
malheureux	370	6,17	vieillesse	184	5,26
tectonique	198	6,02	sinistre	182	5,25
industriel	1462	5,93	substance	686	5,24
thromboemboliques	166	5,89	cardiopathie	132	5,21
fâcheux	161	5,64	bouchon	209	5,11
mauvais	515	5,53	hospitalisation	199	5,07
iatrogène	121	5,4	cancer	485	5,04
bénin	124	5,39	vol	776	5,01
médicamenteux	152	5,3	collision	155	5,0
médian	151	5,29	noyade	119	5,0
malencontreux	110	5,23	insuffisance	233	4,97
mar	99	5,13	tsunami	141	4,94
ischmique	92	5,04	blessé	227	4,9
radiologique	106	4,98	dommage	594	4,69

iovasculaires	89	4,96	myo	90	4,69
responsable	508	4,93	négligence	120	4,67
évitable	87	4,86	accrochage	101	4,63
aigu	136	4,83	absentéisme	92	4,56
allergique	120	4,81	brûlure	111	4,41
traumatique	91	4,79	coma	88	4,36
ischmiques	73	4,71	infraction	179	4,28
géologique	106	4,58	rapatriement	71	4,11
chimique	311	4,58	bagarre	84	4,11
			intoxication	80	4,1
			imprévu	67	4,09
			assurance-maladie	64	4,08
			déversement	64	4,05
			assureur	97	4,01
			vandalisme	63	3,99
			divorce	120	3,95
			mort	884	3,92
			bri|bris	56	3,91
			drame	139	3,87
objet_de	90675	15,33	sujet_de	83428	14,1
----------			----------		
survenir	5897	8,78	survenir	8315	9,29
simuler	248	5,6	impliquer	2024	5,92
recenser	326	5,58	dûs	120	5,44
Un	913	4,99	recenser	169	4,69
lier	1874	4,79	voiturer	75	4,01
frôler	141	4,78	simuler	71	3,86
signaler	403	4,74	multiplier	161	3,78
dénombrer	95	4,64	endeuiller	29	3,27
relater	157	4,53	coûter	183	3,23
eviter	63	4,24	imputer	36	3,23

Depuis	130	4,17	incomber	32	3,23
enregistrer	496	4,07	bouleverser	58	3,04
advenir	65	4,03	advenir	30	2,99
minimiser	82	3,75	indemniser	30	2,99
victimer	49	3,71	médiatiser	33	2,9
enquêter	59	3,67	couté	22	2,88
d	556	3,39	provenir	111	2,85
prvenir	31	3,36	CFIT	17	2,74
comptabiliser	48	3,35	endommager	51	2,63
indemniser	37	3,21	alléguer	21	2,62
imputer	37	3,19	faillir	33	2,6
commettre	161	3,15	accroire	33	2,51
médiatiser	41	3,15	redouter	33	2,5
Eviter	26	3,07	graver	39	2,36
prétexter	27	3,01	relier	84	2,35
u	44	2,93	bpa	13	2,35
cét	21	2,91	préméditer	15	2,27
remémorer	28	2,89	chuter	28	2,19
etait	122	2,81	indéterminer	17	2,17
commémorer	28	2,76	relater	29	2,14
rapporter	121	2,68	élucider	15	2,11
réparer	63	2,57	conduire	138	2,08
			paralyser	18	1,97
			subvenir	11	1,89
pp_de	80724	13,64	pp_du	25923	4,38
----------			----------		
voiture	23748	8,98	travail	20234	6,58
Tchernobyl	1514	8,87	travail-maladies	60	6,23
automobile	1710	8,11	Concorde	76	6,11
décompression	503	7,48	Rio-Paris	31	5,25
circulation	2106	7,11	Boeing	51	5,13
parcours	2727	6,89	Monges	28	5,12
hélicoptère	631	6,86	travail-maladie	20	4,66

trajet	890	6,8	Concordia	14	3,76
plongée	552	6,57	trajet	83	3,73
Three	236	6,49	tunnel	57	3,73
criticité	214	6,39	Bruziers	9	3,51
chasse	924	6,2	réacteur	41	3,42
roulage	208	6,16	Torrey	8	3,3
parapente	181	5,92	Costa	19	3,29
autocar	176	5,73	Travail	37	3,23
bus	417	5,18	Koursk	8	3,23
bagnole	95	4,89	relief	49	3,18
avalanche	91	4,82	travail-	7	3,12
Columbia	81	4,78	camionnage	8	3,1
dimensionnement	79	4,77	Niagara	9	3,07
Felipe	76	4,76	motard	21	2,95
motocyclette	70	4,67	Mont	23	2,88
hydravion	69	4,67	bus	71	2,85
AZF	64	4,63	pétrolier	9	2,83
noyade	60	4,38	prestige	26	2,78
rallye	94	4,33	téléphérique	7	2,74
Bonneval	49	4,26	deux-roues	8	2,72
tracteur	67	4,22	Zeppelin	6	2,71
motoneige	53	4,22	Braer	5	2,66
anesthésie	72	4,1	Kalitta	5	2,65
Roswell	45	4,08	commissariat	17	2,61
piéton	67	4,07	dirigeable	6	2,61
ascenseur	85	4,05	GP	10	2,58
préservatif	65	4,02	Hindenburg	5	2,57
Buizingen	40	4,0	travails	5	2,52
jet-ski	40	3,98	Tunnel	5	2,5
plain-pied	39	3,83	Tricastin	5	2,49
ULM	37	3,8	Gothard	5	2,44
Marcoule	35	3,79	Titanic	7	2,44
Challenger	35	3,75	BST	6	2,44
irradiation	40	3,72	ferry	7	2,39
plonge	35	3,67	anticoagulant	5	2,34

pipi	46	3,65	enfantement	5	2,34
Kubica	33	3,62	Tupolev	4	2,3
poussette	44	3,61	Mont-Blanc	6	2,28
motocross	32	3,6	motocycliste	5	2,28
TMI	30	3,58	Cougar	4	2,23
surdosage	31	3,51	Mugello	4	2,22
montagne	211	3,48	Mans	13	2,1
Courbons	27	3,45	F-BTSC	3	1,92
Seveso	27	3,42	Drayères	3	1,92
Ailefroide	26	3,38	Gondran	3	1,91
balai	42	3,37	Transrapid	3	1,9
contraception	36	3,3	trava	3	1,9
Swissair	25	3,29	golfe	9	1,89
alpinisme	26	3,23	fourgon	5	1,89
tramway	40	3,18	TWA	3	1,88
SR	26	3,18	tokamak	3	1,87
charrette	28	3,17	camion-citerne	3	1,86
Robert	77	3,14	sous-marin	8	1,82
tondeur|tondeuse	27	3,12	viaduc	4	1,78
Windscale	21	3,09	août	59	1,76
dcompression	21	3,08	concorde	4	1,73
			Audi	6	1,72
			Challenger	3	1,71
			Drac	3	1,71
			col	24	1,67
			rafale	7	1,65
			camion	25	1,63
			constellation	5	1,62
			PN	3	1,59
			pain	46	1,56
			quotidien	43	1,52
pp_en	4283	0,72	adj_sujet_of	3855	0,65

rentrant	22	5,33	imputable	74	6,96
Hongrie	48	4,58	évitable	31	6,56
deux-roues	10	4,18	fréquent	293	6,21
canyon	15	3,98	inévitable	111	5,88
guadeloupe	4	3,92	rarissime	22	5,86
kite	4	3,81	bénin	10	5,06
tort	84	3,62	rare	298	4,85
rallye	23	3,29	arrivé	16	4,75
merde	51	3,25	imprévisible	30	4,38
Formule	6	2,91	regrettable	11	4,3
bref	3	2,74	probable	35	4,0
apnée	5	2,7	négligeable	14	3,9
intersection	5	2,69	grave	170	3,81
mer	164	2,62	infinitésimal	3	3,57
EPS	4	2,59	gravissime	3	3,51
roller	4	2,56	nébuleux	3	3,2
moto	57	2,55	proportionnel	13	3,14
pagaille	4	2,4	flou	15	3,08
sortant	4	2,27	accidentel	12	3,07
dehors	46	2,12	envisageable	10	2,98
coiffe	3	2,11	révélateur	9	2,91
direct	14	2,08	inclus	9	2,85
Manche	3	1,95	vite	3	2,81
aout	6	1,91	impensable	3	2,8
plongée	13	1,9	passible	3	2,63
inclusion	5	1,86	préoccupant	4	2,6
aviation	9	1,79	bête	9	2,29
Ontario	9	1,65	moindre	16	2,28
chaîne	50	1,62	stressant	3	2,09
autocar	3	1,59	omniprésent	6	2,03
Californie	8	1,35	compréhensible	5	1,85

psychiatrie	4	1,28	comparable	10	1,54
Suisse	27	1,24	primordial	8	1,35
gymnastique	3	1,15	postérieur	4	1,29
République	3	1,1	obligatoire	20	1,24
roue	21	1,06	sujet	8	1,18
cross	3	0,93	synonyme	3	1,11
carrefour	5	0,91	titulaire	4	0,81
qualification	14	0,89	suspect	3	0,65
milieu	98	0,85	responsable	22	0,64
telechargement	4	0,81	consécutif	6	0,25
général	22	0,77	originaire	3	0,01
hausse	14	0,73	caractéristique	3	-0,02
interne	4	0,62	mineur	4	-0,07
vertu	17	0,61			
amont	4	0,6			
vol	31	0,56			
marge	10	0,5			
escalade	3	0,4			
F	5	0,37			
gare	7	0,22			
Libye	4	0,17			
février	20	0,01			
meurtre	6	-0,09			
pp_au	3600	0,61	prédicat	2221	0,38
----------			----------		
Nurburgring	6	5,42	shoah	3	4,41
coiffe	28	5,38	Nombreux	3	3,48
Mugello	6	5,13	CAN	5	3,25
décollage	26	4,1	Tchernobyl	6	2,82
volant	63	3,91	Malawi	3	2,79
Bol	3	3,81	noyade	3	2,64
intersection	10	3,73	crash	4	1,99

GP	12	3,59	illumination	4	1,9
carrefour	23	3,12	accident	58	1,89
reliure	6	3,09	naufrage	4	1,8
BST	3	2,82	Quels	3	1,49
rallye	16	2,78	occident	3	1,29
passage	198	2,6	C'	166	1,14
navire	74	2,6	Vichy	3	0,8
Japon	66	2,51	n	8	0,66
vertèbre	4	2,47	Valence	3	0,66
verso	3	2,42	Coupe	4	0,61
abord	24	2,41	défaite	9	0,54
rond-point	3	2,36	décès	11	0,5
Mans	9	1,94	premier|première	6	0,46
virage	13	1,81	grossesse	10	0,23
ralenti	5	1,8	explosion	6	0,19
Grand	30	1,65	hier	5	0,17
japon	3	1,59	Belgique	14	0,11
Cambodge	5	1,58	trajet	6	0,11
genou	25	1,52	baiser	5	-0,27
conséquence	77	1,44	massacre	3	-0,58
ski	16	0,96	croix	3	-1,13
Texas	4	0,91	moto	4	-1,27
tunnel	6	0,73	péché	3	-1,49
Caire	4	0,69	substance	5	-1,57
pli	4	0,55	second	3	-1,66
automobiliste	3	0,43			
manoir	3	0,43			
total	7	0,31			
moulin	4	0,24			
compteur	4	-0,05			
rugby	3	-0,16			
laboratoire	10	-0,3			
domicile	13	-0,63			
quotidien	9	-0,67			
stand	3	-0,68			
boulot	6	-0,71			
pont	7	-0,82			
lac	5	-0,88			

tort	3	-1,18			
maximum	5	-1,26			
e	9	-1,78			
pp_à	1964	0,33	pp_avec	1627	0,28
----------			----------		
Talladega	5	6,11	désincarcération	5	5,92
cinétique	12	6,03	arrt	14	5,7
Boumnyebel	3	5,59	blessé	167	5,26
Monza	10	5,36	délit	94	4,44
répétition	126	4,66	Leanna	3	4,41
Indianapolis	6	4,3	alcoolémie	8	4,15
cyclomoteur	4	4,01	Jenson	3	3,89
Signes	3	3,96	arrêt	208	3,75
Daytona	3	3,92	incapacité	23	3,15
Three	3	3,78	lésion	30	2,78
Fukushima	14	3,65	Jasper	3	2,69
Roswell	3	3,52	blessure	74	2,65
Disneyland	3	3,0	retournement	5	2,35
Monaco	20	2,74	Lewis	5	1,99
retardement	3	2,63	ivresse	4	1,82
gogo	3	2,39	interruption	7	1,61
Hz	3	2,33	collision	4	1,35
bord	129	2,26	alcool	29	1,25
Compiègne	3	2,17	décès	14	0,85
Libreville	3	2,12	tiers	14	0,83
Beijing	4	1,94	franchise	5	0,4
vélo	43	1,86	déploiement	4	0,25
cause	195	1,68	responsabilité	29	0,02
Phoenix	3	1,65	marchandise	10	-0,07

Dublin	3	1,65	fusion	5	-0,09
domicile	32	0,68	autrui	3	-0,64
tort	10	0,56	rejet	3	-0,78
proximité	18	0,3			
Los	3	0,21			
St	7	0,2			
Poudlard	3	0,16			
Montréal	14	0,06			
rejet	5	-0,05			
travers	35	-0,35			
incidence	3	-0,54			
l	5	-1,86			
prédicat_de	1352	0,23	pp_par	1327	0,22
----------			----------		
fatalité	13	3,73	CFIT	8	7,5
suivant|suivante	33	3,72	morsure	60	5,93
résultante	5	3,54	électrisation	3	5,59
légion	14	2,36	strangulation	4	5,07
accident	58	1,89	glissade	12	4,7
faute	52	1,62	ingestion	11	4,36
surcharge	3	1,19	inadvertance	5	3,77
puisqu	5	1,08	imprudence	4	3,26
suivant	4	0,55	écrasement	3	2,99
bienvenu	3	0,13	avalanche	4	2,33
conséquence	29	0,04	inhalation	3	2,33
succession	3	-0,41	irradiation	3	2,3
avertissement	3	-0,43	collision	7	2,17
conducteur	3	-0,96	mine	35	1,78
alcool	6	-1,02	distraction	3	1,29
double	3	-1,17	exemple	185	1,19
priorité	6	-1,28	tranche	7	0,67

occasion	14	-1,36	kilomètre	13	0,62
^^	3	-1,67	million	49	0,43
			manque	17	0,16
			catégorie	14	-0,34
			contre	6	-0,49
			excellence	3	-1,14
			province	4	-1,47
pp_sans	505	0,09	pp_sur	454	0,08
----------			----------		
gravité	123	5,09	Mulholland	6	6,86
collision	31	4,33	autoroute	93	4,66
tiers	34	2,11	trois	4	1,27
précédent	18	2,03	accident	23	0,56
blessé	13	1,59	réacteur	4	0,39
signification	9	0,52	chantier	12	0,14
arrêt	12	-0,36	autrui	3	-0,63
ceinture	4	-0,36	Terre	5	-0,68
faute	12	-0,49	l	5	-1,85
permis	5	-0,69			
lendemain	4	-1,33			
pp_entre	277	0,05	pp_pour	260	0,04
----------			----------		
VL	10	6,35	skieur	3	1,66
Vettel	3	2,8	salarié	23	0,32
guillemet	5	2,62	million	23	-0,65
piéton	4	1,39	kilomètre	3	-1,5
camion	13	0,83			
chasseur	3	-1,17			
pp_hors	89	0,02	pp_dans	86	0,01
----------			----------		
dimensionnement	43	6,67	centrale	3	-0,94

intersection	3	2,22			
agglomération	3	-0,31			
pp_depuis	86	0,01	pp_sous	52	0,01
----------			----------		
Tchernobyl	3	1,94	prétexte	3	-0,32
			alcool	4	-1,6
pp_après	49	0,01	pp_contre	24	0,0
----------			----------		
vaccination	13	2,25	obstacle	4	-1,26
accident	4	-1,96			

Annexe 2 : les 101 cliques de l'unité lexicale « accident »

accident, adversité, coup du sort, malchance, malheur, tribulation
accident, affaire, aventure, cas, évènement, fait
accident, affaire, aventure, épisode, évènement, fait
accident, affaire, cas, circonstance, évènement, fait
accident, affaire, cas, circonstance, évènement, occasion
accident, affaire, circonstance, épisode, évènement, fait
accident, avatar, aventure, évènement, incident, mésaventure
accident, avatar, aventure, évènement, incident, péripétie
accident, aventure, cas, évènement, fait, incident
accident, aventure, épisode, évènement, fait, incident
accident, aventure, épisode, évènement, incident, péripétie
accident, calamité, cataclysme, catastrophe, malheur, tribulation
accident, calamité, cataclysme, catastrophe, évènement, malheur
accident, cas, circonstance, évènement, fait, incident
accident, cas, circonstance, évènement, hasard, occasion
accident, cas, circonstance, évènement, incident, occasion
accident, circonstance, épisode, évènement, fait, incident
accident, circonstance, épisode, évènement, incident, péripétie
accident, accroc, anicroche, complication, contretemps
accident, accroc, anicroche, contretemps, incident
accident, adversité, calamité, malheur, tribulation
accident, affaire, aventure, évènement, mésaventure
accident, aléa, aventure, évènement, hasard
accident, aléa, évènement, hasard, vicissitude
accident, anicroche, complication, contretemps, ennui
accident, anicroche, contretemps, ennui, incident
accident, anicroche, ennui, incident, pépin
accident, avatar, évènement, mésaventure, vicissitude
accident, aventure, cas, évènement, hasard
accident, calamité, chagrin, malheur, tribulation
accident, catastrophe, drame, évènement, tragédie
accident, catastrophe, malheur, revers, tribulation
accident, chagrin, ennui, malheur, tribulation
accident, circonstance, incidence, incident, occasion
accident, coup du sort, hasard, malchance, tribulation
accident, malchance, malheur, mésaventure, tribulation
accident, malchance, malheur, mésaventure, tuile
accident, malchance, malheur, revers, tribulation
accident, malchance, mésaventure, pépin, tribulation
accident, malchance, mésaventure, pépin, tuile

accident, malheur, méchef, mésaventure, tribulation
à-coup, accident, accroc, anicroche, complication
à-coup, accident, anicroche, complication, ennui
à-coup, accident, anicroche, ennui, pépin
accident, accroc, incident, panne
accident, accrochage, affaire, choc
accident, accrochage, affaire, complication
accident, accrochage, choc, collision
accident, affaire, complication, ennui
accident, aspérité, inégalité, irrégularité
accident, avatar, ennui, incident
accident, avatar, ennui, tuile
accident, avatar, mésaventure, tuile
accident, aventure, hasard, tribulation
accident, aventure, mésaventure, tribulation
accident, aventure, revers, tribulation
accident, catastrophe, coup, malheur
accident, catastrophe, évènement, péripétie
accident, choc, revers, vicissitude
accident, conséquence, effet, résultat
accident, contretemps, évènement, incident
accident, effet, évènement, résultat
accident, ennui, malheur, tuile
accident, ennui, pépin, tribulation
accident, ennui, pépin, tuile
accident, hasard, malchance, manque de pot
accident, hasard, tribulation, vicissitude
accident, incident, mésaventure, pépin
accident, malencontre, malheur, mésaventure
accident, mésaventure, tribulation, vicissitude
accident, pli, plissement, repli
accident, revers, tribulation, vicissitude
accident, évènement, malheur, mésaventure
accident, accrochage, incident
accident, adversité, avatar
accident, adversité, circonstance
accident, aléa, ennui
accident, bûche, chute
accident, choc, coup
accident, choc, effet
accident, choc, explosion
accident, chute, mouvement
accident, contretemps, malchance
accident, coup, hasard

accident, dénivellation, inégalité
accident, dérapage, tête-à-queue
accident, effet, incidence
accident, fait, phénomène
accident, mouvement, relief
accident, plissement, relief
accident, repli, revers
à-coup, accident, choc
accident, artéfact
accident, attribut
accident, capotage
accident, coup dur
accident, embardée
accident, lésion
accident, montagne
accident, mouvement de terrain
accident, tonneau

Annexe 3 : les sources validées à propos de l'accident de *Deepwater Horizon*

1 Achenbach, J., 2011. A hole at the bottom of the sea: the race to kill the BP oil gusher, 1. Simon & Schuster hardcover ed. ed. Simon & Schuster, New York.
2 Alexander, K., 2010. The 2010 Oil Spill: Minerals Management Service (MMS) and National Environmental Policy Act (NEPA). DTIC Document, p. 28.
3 Allen, Thad W., 2010. National Incident Commander's Report: MC252 Deepwater Horizon.
4 Bly, M., BP, 2010. Deepwater Horizon accident investigation report. Diane Publishing.
5 BP, 2010. Deepwater Horizon Containment and Response: Harnessing Capabilities and Lessons Learned.
6 Cameron, J., 2010. Considering Technical Options for Controlling the BP Blowout in the Gulf of Mexico.
7 Camilli, R., Di Iorio, D., Bowen, A., Reddy, C.M., Techet, A.H., Yoerger, D.R., Whitcomb, L.L., Seewald, J.S., Sylva, S.P., Fenwick, J., 2012. Acoustic measurement of the Deepwater Horizon Macondo well flow rate. Proc. Natl. Acad. Sci. U. S. A. 109, 6. https://doi.org/10.1073/pnas.1100385108
8 Deepwater Horizon Study Group, 2011. Final Report on the Investigation of the Macondo Well Blowout Deepwater Horizon Study Group.
9 Det Norske Veritas, 2011. Final Report For United States Department Of The Interior Bureau Of Ocean Energy Management, Regulation, And Enforcement, Forensic Examination Of Deepwater Horizon Blowout Preventer.
10 Eude, T., Napoli, A., Guarnieri, F., 2016. A thorough analysis of the engineering solutions deployed to stop the oil spill following the deepwater horizon disaster. Chemical Engineering Transactions 6. https://doi.org/10.3303/CET1648130
11 Gulf Coast Ecosystem Restoration Council, 2014. FY2014 Annual Report to Congress Gulf Coast Ecosystem Restoration Council.
12 Hagerty, C.L., 2010. Deepwater Horizon oil spill: Selected issues for Congress. DIANE Publishing.
13 Heidi Avery, 2010. The Ongoing Administration-Wide Response to the Deepwater BP Oil Spill _ The White House [WWW Document].
14 Hickman, S.H., Hsieh, P.A., Mooney, W.D., Enomoto, C.B., Nelson, P.H., Mayer, L.A., Weber, T.C., Moran, K., Flemings, P.B., McNutt, M.K., 2012. Scientific basis for safely shutting in the Macondo Well after the April 20, 2010 Deepwater Horizon blowout. PNAS 109, 6. https://doi.org/10.1073/pnas.1115847109
15 Hsieh, P., 2010. Computer simulation of reservoir depletion and oil flow from the Macondo well following the Deepwater Horizon blowout.
16 Judge Barbier, 2015a. Findings Of Fact And Conclusions Of Law Penalty Phase.
17 Judge Barbier, 2015b. Findings Of Fact And Conclusions Of Law Phase Two Trial.
18 Judge Barbier, 2014. Findings Of Fact And Conclusions Of Law Phase One Trial.
19 Keenan, P., Direction Of Commander, N.S.S.C., 2011. U.S. Navy Salvage Report Deepwater Horizon Oil Spill Response.

20 King, R.O., 2010. Deepwater horizon oil spill disaster: Risk, recovery, and insurance implications 24.
21 Lee, Y.-G., Garza-Gomez, X., 2012. Total cost of the 2010 deepwater horizon oil spill reflected in US stock market. Journal of Accounting and Finance 12, 11.
22 Lubchenco, J., McNutt, M.K., Dreyfus, G., Murawski, S.A., Kennedy, D.M., Anastas, P.T., Chu, S., Hunter, T., 2012. Science in support of the Deepwater Horizon response. PNAS 109, 10. https://doi.org/10.1073/pnas.1204729109
23 McNutt, M.K., Camilli, R., Crone, T.J., Guthrie, G.D., Hsieh, P.A., Ryerson, T.B., Savas, O., Shaffer, F., 2012a. Review of flow rate estimates of the Deepwater Horizon oil spill. PNAS 109, 8. https://doi.org/10.1073/pnas.1112139108
24 McNutt, M.K., Chu, S., Lubchenco, J., Hunter, T., Dreyfus, G., Murawski, S.A., Kennedy, D.M., 2012b. Applications of science and engineering to quantify and control the Deepwater Horizon oil spill. PNAS 109, 8. https://doi.org/10.1073/pnas.1214389109
25 MDL - 2179 Oil Spill by the Oil Rig "Deepwater Horizon" | Eastern District of Louisiana | United States District Court [WWW Document], n.d. URL http://www.laed.uscourts.gov/OilSpill/OilSpill.htm (accessed 5.13.18).
26 MDL 2179 Trial Docs - Phase One [WWW Document], n.d. URL http://www.mdl2179trialdocs.com/index.php?page=phase1 (accessed 7.31.18).
27 MDL 2179 Trial Docs - Phase Three [WWW Document], n.d. URL http://www.mdl2179trialdocs.com/index.php?page=phase3 (accessed 7.31.18).
28 MDL 2179 Trial Docs - Phase Two [WWW Document], n.d. URL http://www.mdl2179trialdocs.com/index.php?page=phase2 (accessed 7.31.18).
29 National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, 2011a. Stopping The Spill: The Five-Month Effort To Kill The Macondo Well ---Draft--- Staff Working Paper No. 6.
30 National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, 2011b. Response and clean-up technology research and development and the BP Deepwater Horizon oil spill.
31 National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, 2011c. Industry's Role in Supporting HSE Standards: Options and Models for the Offshore Oil and Gas Sector Staff Working Paper No. 9.
32 National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, 2011d. Decision-Making Within The Unified Command Staff Working Paper No. 2.
33 National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, 2011e. Deep water: the Gulf oil disaster and the future of offshore drilling : report to the President: recommendations. National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling : For sale by the Supt. of Docs., U.S. G.P.O., [Washington, D.C.].
34 National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, 2011f. Deep water: the Gulf oil disaster and the future of offshore drilling : report to the President. National Commission on the BP Deepwater Horizon Oil

Spill and Offshore Drilling : For sale by the Supt. of Docs., U.S. G.P.O., [Washington, D.C.].
35 National Incident Command, Interagency Solutions Group, Flow Rate Technical Group, 2011. Assessment of Flow Rate Estimates for the Deepwater Horizon / Macondo Well Oil Spill.
36 Nixon, Z., Zengel, S., Baker, M., Steinhoff, M., Fricano, G., Rouhani, S., Michel, J., 2016. Shoreline oiling from the Deepwater Horizon oil spill. Marine Pollution Bulletin 9. https://doi.org/10.1016/j.marpolbul.2016.04.003
37 Occupational Safety and Health Administration, 2011. Deepwater Horizon Oil Spill: OSHA's Role in the Response.
38 Office of the Maritime Administrator, 2011. Republic of the Marshall Islands Deepwater Horizon Marine Casualty Investigation Report.
39 Paul Hsieh, 2016. transcript--TE-PH.pdf.
40 Ramseur, J.L., Hagerty, C.L., 2013. Deepwater Horizon oil spill: Recent activities and ongoing developments. Congressional Research Service. January 31, 21.
41 Steven Chu, 2016. transcript--TE-SC.pdf.
42 Sylves, R.T., Comfort, L.K., 2012. The Exxon Valdez and BP Deepwater Horizon Oil Spills Reducing Risk in Socio-Technical Systems. American Behavioral Scientist 56, 27.
43 Tableau Public [WWW Document], n.d. . Tableau Public. URL https://public.tableau.com/en-us/s/ (accessed 7.27.18).
44 The Bureau Of Ocean Energy Management, Regulation And Enforcement, 2011. Report Regarding The Causes Of The April 20, 2010 Macondo Well Blowout.
45 The Bureau Of Ocean Energy Management, Regulation And Enforcement, 2010. Increased Safety Measures For Energy Development On The Outer Continental Shelf.
46 The Federal Interagency Solutions Group, 2010. Deepwater Horizon oil budget calculator.
47 United States Coast Guard, 2011a. BP Deepwater Horizon Oil Spill Incident Specific Preparedness Review (ISPR).
48 United States Coast Guard, 2011b. United States Coast Guard Report of Investigation into the Circumstances Surrounding the Explosion, Fire, Sinking and Loss of Eleven Crew Members Aboard the Mobile Offshore Drilling Unit Deepwater Horizon In The Gulf Of Mexico April 20 – 22, 2010 (No. MISLE Activity Number: 3721503).
49 United States Coast Guard, 2011c. Federal On Scene Coordinator Report to the National Response Team.
50 US Chemical Safety And Hazard Investigation Board, 2014. Investigation Report Explosion And Fire At The Macondo Well.
51 U.S. Environmental Protection Agency Office Of Inspector General, 2011. Revisions Needed to National Contingency Plan Based on Deepwater Horizon Oil Spill U.S. Environmental Protection Agency Office Of Inspector General.
52 Westerholm, D., 2016. transcript--TE-DW.pdf.

Annexe 4 : les données issues de KH coder à propos de la science sur *Deepwater Horizon*

1. Le tableau de co-occurrences avec « *accident* »

1	major	Adj	105	100	5	2	2	0	11	85	0	0	3	0	2	47,332
2	precursor	Noun	11	3	8	2	1	0	0	0	8	0	0	0	0	18,631
3	causation	Noun	8	1	7	0	1	0	0	0	7	0	0	0	0	15,631
4	prevention	Noun	16	3	13	1	0	1	1	0	11	1	0	0	1	15,136
5	investigation	Noun	19	3	16	1	0	1	1	0	11	1	3	0	1	15,039
6	Amyotte20161	ProperNoun	2	1	1	1	0	0	0	0	0	0	0	0	1	13,823
7	FUKUSHIMA	ProperNoun	10	5	5	0	0	0	1	4	0	2	0	2	1	12,44
8	fatal	Adj	4	4	0	0	1	2	0	1	0	0	0	0	0	12,172
9	terrorist	Adj	4	0	4	0	0	0	0	0	0	1	2	1	0	12,172
10	frequent	Adj	7	6	1	0	1	2	0	3	0	0	0	1	0	12,067
11	maintenance-related	Adj	3	2	1	0	0	0	2	0	0	1	0	0	0	11,847
12	occur	Verb	21	1	20	0	1	0	0	0	6	6	4	2	2	11,122
13	severe	Adj	8	8	0	0	0	0	0	8	0	0	0	0	0	10,962
14	serious	Adj	9	6	3	1	0	0	1	4	0	1	0	0	2	10,301
15	severity	Noun	6	6	0	0	1	3	2	0	0	0	0	0	0	10,264
16	Port	ProperNoun	3	3	0	0	0	1	0	2	0	0	0	0	0	10,188
17	legislation	Noun	7	1	6	1	0	0	0	0	0	2	0	1	3	10,011
18	catastrophic	Adj	8	7	1	0	0	0	0	7	0	0	1	0	0	9,8
19	atypical	Adj	2	2	0	0	0	0	0	2	0	0	0	0	0	9,673
20	barrier-based	Adj	2	2	0	0	0	0	0	2	0	0	0	0	0	9,673
21	Dai-ichi	ProperNoun	2	0	2	0	0	0	0	0	0	0	0	0	2	9,673
22	fear	Verb	2	2	0	0	0	0	0	2	0	0	0	0	0	9,673
23	matter	Verb	2	1	1	0	0	1	0	0	1	0	0	0	0	9,673
24	non-environmental	Adj	2	2	0	0	0	0	0	2	0	0	0	0	0	9,673
25	non-fatal	Adj	2	2	0	0	0	0	0	2	0	0	0	0	0	9,673
26	Reputational	ProperNoun	2	2	0	0	2	0	0	0	0	0	0	0	0	9,673
27	sediment	Verb	2	1	1	1	0	0	0	0	1	0	0	0	0	9,673
28	Severe	ProperNoun	2	2	0	0	0	0	0	2	0	0	0	0	0	9,673
29	strike	Verb	2	1	1	0	1	0	0	0	1	0	0	0	0	9,673
30	Dalian	ProperNoun	3	3	0	1	0	2	0	0	0	0	0	0	0	9,049
31	scenario	Noun	14	0	14	0	0	0	0	0	10	0	0	2	2	8,948
32	report	Noun	10	2	8	0	0	1	1	0	7	0	1	0	0	8,864
33	contributor	Noun	5	1	4	0	1	0	0	0	4	0	0	0	0	8,461
34	cost	Noun	12	6	6	1	0	2	3	0	1	1	0	2	2	8,294
35	prevent	Verb	9	8	1	0	0	2	5	1	0	0	0	1	0	8,08
36	nuclear	Adj	12	12	0	2	1	0	1	8	0	0	0	0	0	8,023
37	industrial	Adj	12	10	2	0	2	1	0	7	0	1	1	0	0	7,958
38	most	Adv	13	8	5	1	2	1	4	0	0	0	3	0	2	7,864
39	bigger	Adj	2	2	0	0	1	1	0	0	0	0	0	0	0	7,815
40	dalian	Adj	2	2	0	0	0	2	0	0	0	0	0	0	0	7,815
41	hydroelectric	Adj	2	0	2	0	0	0	0	0	0	1	0	1	0	7,815
42	MATA-D	ProperNoun	2	0	2	0	0	0	0	0	0	0	0	1	1	7,815
43	Penglai	ProperNoun	2	2	0	0	1	0	1	0	0	0	0	0	0	7,815

44	Deepwater	ProperNoun	20	15	5	2	1	0	12	0	0	0	2	2	1	7,736
45	cause	Noun	11	4	7	0	3	1	0	0	2	0	0	3	2	7,631
46	Horizon	ProperNoun	19	15	4	0	2	1	0	12	0	0	0	2	2	7,331
47	predictive	Adj	5	2	3	0	0	0	0	2	0	0	1	0	2	7,167
48	offshore	Adj	18	8	10	0	1	2	1	4	0	1	6	2	1	7,004
49	comparative	Adj	4	3	1	2	0	1	0	0	0	0	0	1	0	6,93
50	1980	Noun	1	0	1	0	0	0	0	0	0	0	0	1	0	6,84
51	AcostaGonzalez201624	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	6,84
52	Aim	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	6,84
53	Albrechtsen	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
54	Albrechtsen201584	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
55	Amercentrale	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	6,84
56	appraise	Verb	1	0	1	0	0	0	0	0	0	0	0	0	1	6,84
57	area/_/access	Noun	1	0	1	0	0	0	0	0	0	0	1	0	0	6,84
58	Arstad2017114	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
59	auto	Noun	1	1	0	0	0	0	0	1	0	0	0	0	0	6,84
60	Baia	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	6,84
61	beset	Verb	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
62	biggest	Adj	1	1	0	0	0	1	0	0	0	0	0	0	0	6,84
63	bugged	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	6,84
64	categorise	Verb	1	1	0	0	0	0	0	1	0	0	0	0	0	6,84
65	centralized	Adj	1	0	1	0	0	0	0	0	0	1	0	0	0	6,84
66	Challenger	ProperNoun	1	1	0	0	0	1	0	0	0	0	0	0	0	6,84
67	cleveland201485	Noun	1	0	1	0	0	0	0	0	0	0	0	0	1	6,84
68	contemplate	Verb	1	0	1	0	0	0	0	0	0	0	1	0	0	6,84
69	corollary	Noun	1	0	1	0	0	0	0	0	0	0	0	1	0	6,84
70	dataset	Verb	1	0	1	0	0	0	0	0	1	0	0	0	0	6,84
71	dent	Noun	1	0	1	0	0	0	0	0	0	0	0	1	0	6,84
72	Disproportion	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	6,84
73	Dong2017	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	6,84
74	dragon	Verb	1	0	1	0	0	0	0	0	0	1	0	0	0	6,84
75	dutch	Adj	1	1	0	0	0	1	0	0	0	0	0	0	0	6,84
76	electro-mechanical	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	6,84
77	elevate	Verb	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
78	elgheriani201752	Noun	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
79	evidently	Adv	1	1	0	0	0	0	1	0	0	0	0	0	0	6,84
80	Ferdous	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
81	fractional	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	6,84
82	Fund	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
83	hayworth20122005	Noun	1	0	1	0	0	0	0	0	0	0	0	0	1	6,84
84	Hurme	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
85	injuries/_/death	Noun	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
86	inspiration	Noun	1	0	1	0	0	0	0	0	0	0	1	0	0	6,84
87	institutionalize	Verb	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
88	interior	Noun	1	0	1	0	0	0	0	0	0	0	0	0	1	6,84
89	ismail201418	Noun	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
90	Kidam201361	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	6,84
91	Krimsk	ProperNoun	1	1	0	0	0	0	0	1	0	0	0	0	0	6,84

92	Laine	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
93	Laplante	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
94	Mare	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	6,84
95	Markku	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
96	Mearns2017149	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	6,84
97	Moura2017196	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
98	multi-receptor	Noun	1	0	1	0	0	0	0	0	0	0	0	1	0	6,84
99	multi-stressor	Noun	1	0	1	0	0	0	0	0	0	1	0	0	0	6,84
100	non-effect	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	6,84
101	nonstop	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	6,84
102	Normal	ProperNoun	1	1	0	0	0	0	0	1	0	0	0	0	0	6,84
103	Okoh20131060	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
104	overwhelming	Adj	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
105	pasman2015185	Noun	1	0	1	0	0	0	0	0	0	0	0	0	1	6,84
106	Patin	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
107	Pike201737	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	6,84
108	post-disaster	Noun	1	0	1	0	0	0	0	0	0	1	0	0	0	6,84
109	precondition	Noun	1	1	0	0	0	0	1	0	0	0	0	0	0	6,84
110	PSI	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
111	rarer	Noun	1	1	0	0	0	0	0	1	0	0	0	0	0	6,84
112	recorded	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	6,84
113	Refaul	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
114	remark	Noun	1	0	1	0	0	0	0	0	0	1	0	0	0	6,84
115	reniers2014779	Noun	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
116	reportable	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	6,84
117	silva2017319	Noun	1	1	0	0	0	1	0	0	0	0	0	0	0	6,84
118	STAMP	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
119	step-up	Adj	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
120	System-Theoretic	ProperNoun	1	1	0	0	0	0	0	1	0	0	0	0	0	6,84
121	technically	Adv	1	0	1	0	0	0	0	0	0	0	1	0	0	6,84
122	Technological	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	6,84
123	Terrorism	ProperNoun	1	0	1	0	0	0	0	0	0	1	0	0	0	6,84
124	Tutorial	ProperNoun	1	0	1	0	0	0	0	0	0	1	0	0	0	6,84
125	Underwood	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
126	WAP	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	6,84
127	Willey	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
128	Yeosu	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	6,84
129	causality	Noun	2	0	2	0	0	0	0	0	2	0	0	0	0	6,697
130	complacency	Noun	2	1	1	1	0	0	0	0	0	0	0	0	1	6,697
131	credibility	Noun	2	0	2	0	0	0	0	0	2	0	0	0	0	6,697
132	Daiichi	ProperNoun	2	0	2	0	0	0	0	0	0	0	2	0	0	6,697
133	reputational	Adj	2	0	2	0	0	0	0	0	0	0	0	2	0	6,697
134	SHIPP	ProperNoun	2	2	0	1	0	1	0	0	0	0	0	0	0	6,697
135	shortly	Adv	2	1	1	0	0	1	0	0	0	0	0	0	1	6,697
136	undesired	Adj	2	2	0	1	0	1	0	0	0	0	0	0	0	6,697
137	avoid	Verb	4	3	1	0	0	0	2	1	0	0	0	1	0	6,658
138	analysis	Noun	29	12	17	2	3	6	1	0	12	0	1	2	2	6,473
139	occurrence	Noun	6	5	1	1	3	1	0	0	1	0	0	0	0	6,351

140 happen	Verb	4	0	4	0	0	0	0	0	3	1	0	0	0	6,184
141 QRA	ProperNoun	4	1	3	0	0	1	0	0	0	0	0	3	0	6,184
142 aftermath	Noun	3	2	1	0	0	2	0	0	0	0	0	1	0	6,172
143 equipment	Noun	7	5	2	1	0	1	1	2	0	1	0	0	1	6,149
144 fatality	Noun	5	3	2	0	2	0	1	0	0	0	2	0	0	5,976
145 casualty	Noun	2	0	2	0	0	0	0	0	0	1	1	0	0	5,926
146 Chernobyl	ProperNoun	2	2	0	0	0	0	1	1	0	0	0	0	0	5,926
147 dam	Noun	2	0	2	0	0	0	0	0	0	0	1	0	1	5,926
148 energy-related	Adj	2	2	0	0	0	0	2	0	0	0	0	0	0	5,926
149 undergo	Verb	2	0	2	0	0	0	0	0	1	0	1	0	0	5,926
150 cause	Verb	10	4	6	1	1	1	1	0	3	2	0	1	0	5,815
151 hypothetical	Adj	3	2	1	0	0	1	1	0	0	1	0	0	0	5,552
152 update	Verb	4	3	1	1	1	0	1	0	0	0	0	0	1	5,439
153 consequence	Noun	10	7	3	1	1	2	3	0	0	1	1	1	0	5,388
154 Accimaps	ProperNoun	2	2	0	1	0	1	0	0	0	0	0	0	0	5,351
155 prone	Adj	2	2	0	1	0	1	0	0	0	0	0	0	0	5,351
156 follow	Verb	10	7	3	2	0	2	3	0	0	1	1	1	0	5,349
157 differ	Verb	3	0	3	0	0	0	0	0	0	2	1	0	0	5,294
158 BP	ProperNoun	9	7	2	0	6	0	0	1	0	0	0	2	0	5,137
159 potential	Adj	11	9	2	0	2	0	0	7	0	1	0	0	1	5,074
160 Alpha	ProperNoun	3	2	1	0	0	0	0	2	0	0	0	1	0	5,063
161 Piper	ProperNoun	3	2	1	0	0	0	2	0	0	0	1	0	0	5,063
162 fire	Noun	11	10	1	3	2	1	0	4	0	0	0	1	0	4,918
163 attack	Noun	2	0	2	0	0	0	0	0	0	0	0	2	0	4,9
164 symptom	Noun	2	0	2	0	0	0	0	0	0	0	2	0	0	4,9
165 learn	Verb	8	8	0	0	1	2	5	0	0	0	0	0	0	4,828
166 activity-related	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	4,735
167 announcement	Noun	1	1	0	0	0	1	0	0	0	0	0	0	0	4,735
168 aside	Adv	1	0	1	0	0	0	0	0	0	1	0	0	0	4,735
169 Carvalho	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	4,735
170 Columbia	ProperNoun	1	1	0	0	0	0	0	1	0	0	0	0	0	4,735
171 committee	Noun	1	0	1	0	0	0	0	0	0	0	0	1	0	4,735
172 credible	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	4,735
173 Database	ProperNoun	1	0	1	0	0	0	0	0	1	0	0	0	0	4,735
174 depart	Verb	1	0	1	0	0	0	0	0	0	0	0	1	0	4,735
175 downwind	Noun	1	1	0	0	0	1	0	0	0	0	0	0	0	4,735
176 e,g,	Adj	1	0	1	0	0	0	0	0	0	1	0	0	0	4,735
177 earliest	Adj	1	1	0	0	0	1	0	0	0	0	0	0	0	4,735
178 elimination	Noun	1	1	0	0	1	0	0	0	0	0	0	0	0	4,735
179 endeavor	Noun	1	0	1	0	0	0	0	0	0	0	0	1	0	4,735
180 ENSAD	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	4,735
181 exact	Adj	1	1	0	1	0	0	0	0	0	0	0	0	0	4,735
182 Franz	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	4,735
183 front	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	4,735
184 hindsight	Noun	1	1	0	0	0	1	0	0	0	0	0	0	0	4,735
185 i	Noun	1	0	1	0	0	0	0	0	0	0	0	0	1	4,735
186 i&c	Noun	1	0	1	0	0	0	0	0	0	0	0	1	0	4,735
187 ict-based	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	4,735

188 impulsive	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	4,735
189 king	Noun	1	0	1	0	0	0	0	0	0	0	1	0	0	4,735
190 Knoll	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	4,735
191 legitimate	Adj	1	1	0	0	1	0	0	0	0	0	0	0	0	4,735
192 Meiofauna	ProperNoun	1	0	1	0	0	0	0	0	0	1	0	0	0	4,735
193 metric	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	4,735
194 Mile	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	4,735
195 money	Noun	1	0	1	0	0	0	0	0	0	0	1	0	0	4,735
196 multi-attribute	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	4,735
197 natech	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	4,735
198 naval	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	4,735
199 Netherlands	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	4,735
200 orient	Verb	1	0	1	0	0	0	0	0	0	0	0	1	0	4,735

2. Le tableau de co-occurrences de noms propres avec « *accident* »

1 Amyotte20161	ProperNoun	2	1	1	1	0	0	0	0	0	0	0	0	1	13,823
2 FUKUSHIMA	ProperNoun	10	5	5	0	0	0	1	4	0	2	0	2	1	12,44
3 Port	ProperNoun	3	3	0	0	0	1	0	2	0	0	0	0	0	10,188
4 Dai-ichi	ProperNoun	2	0	2	0	0	0	0	0	0	0	0	0	2	9,673
5 Reputational	ProperNoun	2	2	0	0	2	0	0	0	0	0	0	0	0	9,673
6 Severe	ProperNoun	2	2	0	0	0	0	0	2	0	0	0	0	0	9,673
7 Dalian	ProperNoun	3	3	0	1	0	2	0	0	0	0	0	0	0	9,049
8 MATA-D	ProperNoun	2	0	2	0	0	0	0	0	0	0	0	1	1	7,815
9 Penglai	ProperNoun	2	2	0	0	1	0	1	0	0	0	0	0	0	7,815
10 Deepwater	ProperNoun	20	15	5	2	1	0	12	0	0	0	2	2	1	7,736
11 Horizon	ProperNoun	19	15	4	0	2	1	0	12	0	0	0	2	2	7,331
12 AcostaGonzalez201624	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	6,84
13 Aim	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	6,84
14 Albrechtsen	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
15 Albrechtsen201584	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
16 Amercentrale	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	6,84
17 Arstad2017114	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
18 Baia	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	6,84
19 Challenger	ProperNoun	1	1	0	0	0	1	0	0	0	0	0	0	0	6,84
20 Disproportion	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	6,84
21 Dong2017	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	6,84
22 Ferdous	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
23 Fund	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
24 Hurme	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
25 Kidam201361	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	6,84
26 Krimsk	ProperNoun	1	1	0	0	0	0	0	1	0	0	0	0	0	6,84
27 Laine	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
28 Laplante	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
29 Mare	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	6,84
30 Markku	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
31 Mearns2017149	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	6,84
32 Moura2017196	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84

33 Normal	ProperNoun	1	1	0	0	0	0	0	1	0	0	0	0	0	6,84
34 Okoh20131060	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
35 Patin	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
36 Pike201737	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	6,84
37 PSI	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
38 Refaul	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
39 STAMP	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
40 System-Theoretic	ProperNoun	1	1	0	0	0	0	0	1	0	0	0	0	0	6,84
41 Technological	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	6,84
42 Terrorism	ProperNoun	1	0	1	0	0	0	0	0	0	1	0	0	0	6,84
43 Tutorial	ProperNoun	1	0	1	0	0	0	0	0	0	1	0	0	0	6,84
44 Underwood	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	6,84
45 WAP	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	6,84
46 Willey	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	6,84
47 Yeosu	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	6,84
48 Daiichi	ProperNoun	2	0	2	0	0	0	0	0	0	0	2	0	0	6,697
49 SHIPP	ProperNoun	2	2	0	1	0	1	0	0	0	0	0	0	0	6,697
50 QRA	ProperNoun	4	1	3	0	0	1	0	0	0	0	0	3	0	6,184
51 Chernobyl	ProperNoun	2	2	0	0	0	0	1	1	0	0	0	0	0	5,926
52 Accimaps	ProperNoun	2	2	0	1	0	1	0	0	0	0	0	0	0	5,351
53 BP	ProperNoun	9	7	2	0	6	0	0	1	0	0	0	2	0	5,137
54 Alpha	ProperNoun	3	2	1	0	0	0	0	2	0	0	0	1	0	5,063
55 Piper	ProperNoun	3	2	1	0	0	0	2	0	0	0	1	0	0	5,063
56 Carvalho	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	4,735
57 Columbia	ProperNoun	1	1	0	0	0	0	0	1	0	0	0	0	0	4,735
58 Database	ProperNoun	1	0	1	0	0	0	0	0	1	0	0	0	0	4,735
59 ENSAD	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	4,735
60 Franz	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	4,735
61 Knoll	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	4,735
62 Meiofauna	ProperNoun	1	0	1	0	0	0	0	0	0	1	0	0	0	4,735
63 Mile	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	4,735
64 Netherlands	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	4,735
65 Soviet	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	4,735
66 Tokyo	ProperNoun	1	0	1	0	0	0	0	0	0	1	0	0	0	4,735
67 DWH	ProperNoun	6	6	0	0	1	0	4	1	0	0	0	0	0	4,528
68 Learning	ProperNoun	4	4	0	0	0	1	3	0	0	0	0	0	0	4,316
69 Black	ProperNoun	2	0	2	0	0	0	0	0	0	0	0	1	1	4,226
70 WSA	ProperNoun	2	1	1	1	0	0	0	0	0	0	0	0	1	4,226
71 Process	ProperNoun	10	3	7	0	0	0	0	3	1	0	0	2	4	3,875
72 Japan	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	3,784
73 Jet	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	3,784
74 Mumbai	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	3,784
75 PTR-MS	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	3,784
76 RED	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	3,784
77 Testing	ProperNoun	1	0	1	0	0	0	0	0	0	1	0	0	0	3,784
78 Processes	ProperNoun	2	1	1	0	1	0	0	0	0	0	1	0	0	3,536
79 Compensation	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	3,205
80 Contemporary	ProperNoun	1	1	0	0	0	1	0	0	0	0	0	0	0	3,205

81	Hey	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	3,205
82	July	ProperNoun	1	0	1	0	0	0	0	0	0	1	0	0	0	3,205
83	Situation	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	3,205
84	Statistical	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	3,205
85	Awareness	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	2,803
86	Buncefield	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	2,803
87	Crete	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	2,803
88	Internet	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	2,803
89	Investigation	ProperNoun	1	0	1	0	0	0	0	0	1	0	0	0	0	2,803
90	Analysis	ProperNoun	4	1	3	1	0	0	0	0	3	0	0	0	0	2,705
91	Blowout	ProperNoun	3	2	1	1	0	0	0	1	0	1	0	0	0	2,573
92	Safety	ProperNoun	13	4	9	4	0	0	0	0	0	0	0	5	4	2,503
93	Haugen	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	2,5
94	Reniers	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	2,5
95	Prevention	ProperNoun	4	1	3	1	0	0	0	0	0	0	3	0	0	2,42
96	Accident	ProperNoun	3	1	2	1	0	0	0	0	0	0	0	1	1	2,407
97	Culture	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	2,261
98	HMI	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	2,261
99	Hydrocarbon	ProperNoun	2	1	1	1	0	0	0	0	0	0	0	1	0	2,149
100	Amyotte	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	2,064
101	Directive	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	2,064
102	Report	ProperNoun	1	0	1	0	0	0	0	0	0	1	0	0	0	2,064
103	U,S,	ProperNoun	2	0	2	0	0	0	0	0	0	0	2	0	0	2,009
104	China	ProperNoun	2	0	2	0	0	0	0	0	0	2	0	0	0	1,944
105	Eastern	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	1,898
106	OPOL	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	1,898
107	Bruce	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	1,755
108	Island	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	1,755
109	Network	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	1,755
110	Silva	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	1,755
111	Union	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	1,755
112	Hebei	ProperNoun	1	1	0	0	0	0	1	0	0	0	0	0	0	1,631
113	Offshore	ProperNoun	4	3	1	1	0	0	1	1	0	0	0	1	0	1,539
114	Spirit	ProperNoun	1	1	0	0	0	0	0	1	0	0	0	0	0	1,52
115	Theory	ProperNoun	1	0	1	0	0	0	0	0	1	0	0	0	0	1,52
116	Thomson	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	1,52
117	Anderson	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	1,42
118	Applied	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	1,42
119	Drilling	ProperNoun	1	1	0	0	0	0	0	1	0	0	0	0	0	1,42
120	Method	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	1,42
121	FPSO	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	1,331
122	J,R,	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	1,331
123	Paltrinieri	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	1,248
124	North	ProperNoun	2	0	2	0	0	0	0	0	0	0	1	1	0	1,227
125	High	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	1,173
126	New	ProperNoun	2	2	0	0	1	0	1	0	0	0	0	0	0	1,121
127	Fire	ProperNoun	1	1	0	0	0	0	0	1	0	0	0	0	0	1,038
128	Science	ProperNoun	3	0	3	0	0	0	0	0	0	0	0	0	3	0,952

129	Energy	ProperNoun	4	2	2	0	1	0	0	1	0	0	0	2	0	0,914
130	Bay	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	0,868
131	CFD	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	0,868
132	Hans	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	0,868
133	Pasman	ProperNoun	1	1	0	0	1	0	0	0	0	0	0	0	0	0,868
134	Peter	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	0,817
135	Sea	ProperNoun	3	0	3	0	0	0	0	0	0	0	0	1	2	0,811
136	Model	ProperNoun	1	0	1	0	0	0	0	0	1	0	0	0	0	0,769
137	South	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	0,769
138	Texas	ProperNoun	1	0	1	0	0	0	0	0	0	1	0	0	0	0,724
139	Industrial	ProperNoun	1	1	0	0	0	0	0	1	0	0	0	0	0	0,681
140	Nuclear	ProperNoun	1	1	0	0	0	0	0	1	0	0	0	0	0	0,681
141	Mediterranean	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	0,639
142	Paul	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	0,639
143	Europe	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	0,6
144	Macondo	ProperNoun	2	2	0	1	0	0	0	1	0	0	0	0	0	0,451
145	Louisiana	ProperNoun	1	0	1	0	0	0	0	0	0	1	0	0	0	0,393
146	OIL	ProperNoun	6	4	2	2	0	0	2	0	0	0	0	2	0	0,366
147	Khan	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	0,333
148	US	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	0,277
149	States	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	0,25
150	T	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	0,224
151	United	ProperNoun	1	0	1	0	0	0	0	0	0	0	1	0	0	0,037
152	S	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	-0,063
153	Ocean	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	-0,223
154	Risk	ProperNoun	1	1	0	0	0	1	0	0	0	0	0	0	0	-0,562
155	Engineering	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	0	1	-0,622
156	Mexico	ProperNoun	2	0	2	0	0	0	0	0	0	0	0	0	2	-0,76
157	A	ProperNoun	1	1	0	1	0	0	0	0	0	0	0	0	0	-0,795
158	MANAGEMENT	ProperNoun	1	0	1	0	0	0	0	0	1	0	0	0	0	-0,845
159	Journal	ProperNoun	1	0	1	0	0	0	0	0	0	0	0	1	0	-0,93
160	Gulf	ProperNoun	3	0	3	0	0	0	0	0	0	0	2	0	1	-1,016

3. Le tableau des co-occurrences associées avec « *accident* » ET « *model* »

1	predictive	Adj	12	6	6	0	2	0	1	3	3	1	0	2	0	19,318
2	mental	Adj	6	6	0	0	0	0	0	6	0	0	0	0	0	13,66
3	swiss	Adj	4	4	0	0	0	0	4	0	0	0	0	0	0	11,793
4	numerical	Adj	7	4	3	1	0	0	1	2	0	1	0	1	1	11,335
5	linear	Adj	5	2	3	0	0	1	0	1	0	0	1	1	1	11,312
6	epidemiological	Adj	3	2	1	0	0	0	0	2	0	0	0	1	0	10,866
7	barrier-based	Adj	2	2	0	0	0	0	2	0	0	0	0	0	0	10,308
8	multi-plant	Adj	2	2	0	0	0	0	2	0	0	0	0	0	0	10,308
9	sub-surface	Adj	2	1	1	0	0	1	0	0	0	0	0	1	0	10,308
10	probabilistic	Adj	8	8	0	1	0	3	1	3	0	0	0	0	0	9,565
11	hybrid	Adj	3	3	0	0	0	1	0	2	0	0	0	0	0	8,762
12	multiphase	Adj	2	2	0	1	0	0	1	0	0	0	0	0	0	8,339

13 non-linear	Adj	2	1	1	0	0	1	0	0	0	0	0	0	1	8,339
14 proportional	Adj	3	3	0	0	0	0	3	0	0	0	0	0	0	8,061
15 bioanalytical	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	7,289
16 cfd-based	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	7,289
17 compartmented	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	7,289
18 current-generation	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	7,289
19 data-assimilative	Adj	1	1	0	0	1	0	0	0	0	0	0	0	0	7,289
20 delft3d	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	7,289
21 dual-mode	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	7,289
22 fault-tree	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	7,289
23 game-theory	Adj	1	1	0	1	0	0	0	0	0	0	0	0	0	7,289
24 geometric	Adj	1	1	0	1	0	0	0	0	0	0	0	0	0	7,289
25 input-output	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	7,289
26 interpretable	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	7,289
27 management-task-safety	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	7,289
28 met-ocean	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	7,289
29 meta-damage	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	7,289
30 multi-bass	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	7,289
31 particle-tracking	Adj	1	1	0	1	0	0	0	0	0	0	0	0	0	7,289
32 person-to-person	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	7,289
33 physics-based	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	7,289
34 prognostic	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	7,289
35 revised	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	7,289
36 solute-transport	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	7,289
37 steady-state	Adj	1	1	0	0	0	1	0	0	0	0	0	0	0	7,289
38 supra-plant	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	7,289
39 technology-embedded	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	7,289
40 thermodynamic	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	7,289
41 three-level	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	7,289
42 time-invariant	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	7,289
43 transnational	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	7,289
44 trust-repair	Adj	1	1	0	0	1	0	0	0	0	0	0	0	0	7,289
45 two-wave	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	7,289
46 uncored	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	7,289
47 understandable	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	7,289
48 generic	Adj	4	4	0	0	1	0	0	3	0	0	0	0	0	6,885
49 norwegian	Adj	5	4	1	0	0	0	0	4	0	0	0	0	1	6,751
50 corresponding	Adj	3	2	1	0	0	0	2	0	0	0	1	0	0	6,617
51 meteorological	Adj	2	1	1	0	0	0	1	0	0	0	0	1	0	6,339
52 regional	Adj	4	4	0	0	0	0	2	2	0	0	0	0	0	5,699
53 multiple	Adj	6	4	2	0	0	3	0	1	0	0	1	1	0	5,496
54 statistical	Adj	3	3	0	0	1	0	0	2	0	0	0	0	0	5,448
55 applied	Adj	1	1	0	0	0	1	0	0	0	0	0	0	0	5,059
56 asymmetric	Adj	1	1	0	1	0	0	0	0	0	0	0	0	0	5,059
57 concentrated	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	5,059
58 dimensional	Adj	1	1	0	1	0	0	0	0	0	0	0	0	0	5,059

59	diurnal	Adj	1	1	0	0	0	1	0	0	0	0	0	0	0	5,059
60	fifteen	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	5,059
61	nearest	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	5,059
62	one-dimensional	Adj	1	1	0	0	1	0	0	0	0	0	0	0	0	5,059
63	separated	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	5,059
64	spare	Adj	1	0	1	0	0	0	0	0	0	1	0	0	0	5,059
65	testable	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	5,059
66	degraded	Adj	2	2	0	2	0	0	0	0	0	0	0	0	0	4,868
67	deterministic	Adj	2	1	1	0	0	1	0	0	0	0	0	0	1	4,868
68	ecotoxicological	Adj	2	1	1	0	0	0	0	1	0	0	1	0	0	4,868
69	high-resolution	Adj	2	0	2	0	0	0	0	0	0	1	1	0	0	4,868
70	well-known	Adj	2	1	1	0	0	1	0	0	0	0	1	0	0	4,868
71	available	Adj	5	3	2	0	0	3	0	0	0	0	0	1	1	4,63
72	american	Adj	2	2	0	0	1	0	0	1	0	0	0	0	0	4,545
73	imperfect	Adj	3	1	2	0	0	1	0	0	0	2	0	0	0	4,378
74	case-based	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	4,053
75	cross-plant	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	4,053
76	evidence-based	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	4,053
77	graphical	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	4,053
78	hydrodynamic	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	4,053
79	inappropriate	Adj	1	1	0	1	0	0	0	0	0	0	0	0	0	4,053
80	sophisticated	Adj	1	1	0	0	1	0	0	0	0	0	0	0	0	4,053
81	three-step	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	4,053
82	weaker	Adj	1	1	0	1	0	0	0	0	0	0	0	0	0	4,053
83	single	Adj	3	3	0	0	0	0	1	2	0	0	0	0	0	3,886
84	applicable	Adj	2	0	2	0	0	0	0	0	1	1	0	0	0	3,82
85	solid	Adj	2	1	1	0	0	0	1	0	0	0	1	0	0	3,82
86	simple	Adj	3	2	1	0	0	1	1	0	0	0	1	0	0	3,781
87	causal	Adj	2	1	1	0	0	0	0	1	0	0	0	1	0	3,632
88	dynamic	Adj	4	4	0	1	1	0	2	0	0	0	0	0	0	3,465
89	conceptual	Adj	2	2	0	0	0	0	0	2	0	0	0	0	0	3,464
90	234th	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	3,443
91	analytic	Adj	1	1	0	0	0	1	0	0	0	0	0	0	0	3,443
92	former	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	3,443
93	known	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	3,443
94	longer-term	Adj	1	1	0	0	0	1	0	0	0	0	0	0	0	3,443
95	mathematical	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	3,443
96	north	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	3,443
97	redundant	Adj	1	0	1	0	0	0	0	0	0	1	0	0	0	3,443
98	wellbore	Adj	1	0	1	0	0	0	0	0	0	1	0	0	0	3,443
99	capable	Adj	2	0	2	0	0	0	0	0	0	1	1	0	0	3,312
100	different	Adj	10	5	5	1	0	0	1	3	1	1	1	0	2	3,106
101	lagrangian	Adj	2	1	1	0	0	1	0	0	0	0	0	1	0	3,046
102	conditional	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	3,019
103	earlier	Adj	1	0	1	0	0	0	0	0	0	1	0	0	0	3,019
104	explicit	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	3,019

105 exponential	Adj	1	1	0	1	0	0	0	0	0	0	0	0	0	3,019
106 geomorphological	Adj	1	1	0	0	0	1	0	0	0	0	0	0	0	3,019
107 hydrophobic	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	3,019
108 improved	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	3,019
109 sequential	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	3,019
110 three-dimensional	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	3,019
111 accidental	Adj	3	1	2	1	0	0	0	0	0	1	1	0	0	2,968
112 bayesian	Adj	4	2	2	0	0	0	2	0	0	0	0	2	0	2,917
113 real-time	Adj	2	1	1	0	0	0	1	0	0	0	0	0	1	2,718
114 aerial	Adj	1	0	1	0	0	0	0	0	0	1	0	0	0	2,701
115 big	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	2,701
116 circular	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	2,701
117 daily	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	2,701
118 isotopic	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	2,701
119 lowest	Adj	1	0	1	0	0	0	0	0	0	1	0	0	0	2,701
120 parallel	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	2,701
121 stochastic	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	2,701
122 individual	Adj	3	2	1	0	1	0	0	1	0	1	0	0	0	2,564
123 new	Adj	7	6	1	0	1	1	1	3	0	0	1	0	0	2,555
124 additional	Adj	2	0	2	0	0	0	0	0	0	0	2	0	0	2,534
125 independent	Adj	2	1	1	0	0	0	1	0	0	0	0	1	0	2,534
126 common-cause	Adj	1	0	1	0	0	0	0	0	0	1	0	0	0	2,45
127 risky	Adj	1	1	0	0	0	1	0	0	0	0	0	0	0	2,45
128 simulated	Adj	2	0	2	0	0	0	0	0	0	1	0	1	0	2,45
129 computational	Adj	2	1	1	0	0	0	0	1	0	0	0	1	0	2,296
130 previous	Adj	2	1	1	0	0	0	0	1	0	0	0	0	1	2,296
131 systemic	Adj	2	1	1	0	1	0	0	0	0	0	0	1	0	2,296
132 cold-water	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	2,244
133 content	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	2,244
134 following	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	2,244
135 methodological	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	2,244
136 novel	Adj	1	1	0	0	0	1	0	0	0	0	0	0	0	2,244
137 preliminary	Adj	1	1	0	0	1	0	0	0	0	0	0	0	0	2,244
138 general	Adj	3	2	1	0	0	1	1	0	0	1	0	0	0	2,239
139 adaptive	Adj	1	1	0	1	0	0	0	0	0	0	0	0	0	2,071
140 autonomous	Adj	1	1	0	0	1	0	0	0	0	0	0	0	0	2,071
141 bottom	Adj	1	1	0	0	0	1	0	0	0	0	0	0	0	2,071
142 deep-water	Adj	1	1	0	0	1	0	0	0	0	0	0	0	0	2,071
143 partial	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	2,071
144 experimental	Adj	2	0	2	0	0	0	0	0	0	0	0	1	1	2,032
145 able	Adj	2	1	1	0	0	0	1	0	0	1	0	0	0	1,973
146 operational	Adj	4	3	1	0	0	2	0	1	0	0	0	1	0	1,953
147 larger	Adj	1	0	1	0	0	0	0	0	0	1	0	0	0	1,922
148 sandy	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	1,922
149 vertical	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	1,922
150 weak	Adj	1	1	0	1	0	0	0	0	0	0	0	0	0	1,922

151 active	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	1,792
152 local	Adj	2	0	2	0	0	0	0	0	0	1	1	0	0	1,761
153 ecological	Adj	2	1	1	0	0	0	1	0	0	0	0	0	1	1,713
154 mixed	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	1,676
155 oceanographic	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	1,676
156 common	Adj	3	0	3	0	0	0	0	0	1	0	0	0	2	1,59
157 latter	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	1,573
158 preventive	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	1,573
159 acceptable	Adj	1	1	0	0	0	1	0	0	0	0	0	0	0	1,48
160 accurate	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	1,48
161 dominant	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	1,48
162 routine	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	1,48
163 southern	Adj	1	1	0	0	1	0	0	0	0	0	0	0	0	1,48
164 entire	Adj	1	0	1	0	0	0	0	0	0	1	0	0	0	1,395
165 full	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	1,395
166 robust	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	1,395
167 better	Adj	2	0	2	0	0	0	0	0	0	0	2	0	0	1,381
168 analytical	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	1,317
169 atmospheric	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	1,317
170 large-scale	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	1,317
171 physical	Adj	2	2	0	0	0	0	2	0	0	0	0	0	0	1,276
172 coral	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	1,245
173 empirical	Adj	1	1	0	0	0	1	0	0	0	0	0	0	0	1,245
174 significant	Adj	4	3	1	1	0	1	1	0	0	0	0	1	0	1,218
175 further	Adj	2	1	1	0	0	1	0	0	0	1	0	0	0	1,21
176 actual	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	1,178
177 fundamental	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	1,178
178 strong	Adj	1	1	0	0	0	1	0	0	0	0	0	0	0	1,178
179 systematic	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	1,178
180 cognitive	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	1,116
181 distinct	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	1,058
182 internal	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	1,003
183 relative	Adj	1	0	1	0	0	0	0	0	0	0	0	0	1	1,003
184 appropriate	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	0,951
185 extensive	Adj	1	0	1	0	0	0	0	0	0	0	0	1	0	0,951
186 positive	Adj	1	1	0	1	0	0	0	0	0	0	0	0	0	0,951
187 crude	Adj	4	3	1	2	0	1	0	0	0	0	0	1	0	0,946
188 continuous	Adj	1	1	0	0	0	1	0	0	0	0	0	0	0	0,902
189 underwater	Adj	1	1	0	0	0	1	0	0	0	0	0	0	0	0,902
190 wide	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	0,902
191 greater	Adj	1	1	0	1	0	0	0	0	0	0	0	0	0	0,811
192 historical	Adj	1	0	1	0	0	0	0	0	0	1	0	0	0	0,811
193 theoretical	Adj	1	1	0	0	0	0	0	1	0	0	0	0	0	0,811
194 thermal	Adj	1	0	1	0	0	0	0	0	0	0	1	0	0	0,811
195 due	Adj	3	0	3	0	0	0	0	0	2	1	0	0	0	0,81
196 commercial	Adj	1	0	1	0	0	0	0	0	0	1	0	0	0	0,769

197 highest	Adj	1	0	1	0	0	0	0	0	0	1	0	0	0,769
198 economic	Adj	2	1	1	0	0	0	0	1	0	0	0	1	0	0,733
199 efficient	Adj	1	1	0	0	1	0	0	0	0	0	0	0	0	0,729
200 limited	Adj	1	1	0	0	0	0	1	0	0	0	0	0	0	0,729

Annexe 5 : extrait du *staff working paper* n°6, la partie B Cofferdam

Cet extrait comporte 38 phrases :

1. On April 25, as efforts to actuate the BOP stack continued, BP began to consider placing a large containment dome, also known as a cofferdam, over the larger of the two leaks from the broken riser.
2. At the top of the cofferdam, a pipe would channel hydrocarbons to the Discoverer Enterprise, a ship on the surface.
3. Although some initial reports indicated that BP would need as long as four weeks to install the dome, BP was able to move more rapidly.
4. Several cofferdams were already in existence, with BP having used them to recover oil from shallow-water leaks following Hurricanes Katrina and Rita.
5. By May 4, just ten days after first raising the possibility of using a containment dome, BP reported that it had finished modifying for deep-sea use a preexisting dome that was 14 feet wide, 24 feet long, and 40 feet tall.
6. Following an MMS inspection of the Discoverer Enterprise, BP began to lower the 98-ton dome to the sea floor late in the evening of May 6.
7. BP planned to stage a second cofferdam on the sea floor in case the first dam failed.
8. From the beginning, the likelihood of collecting hydrocarbons with the cofferdam was uncertain.
9. Suttles of BP publicly cautioned that a containment dome had only been used successfully in much shallower water.
10. In an interview, he told Commission staff that, according to BP engineers, the chance of success was at best 50 percent.
11. Bob Fryar, a senior BP engineer, warned : this is new technology, it has never been done before.
12. BP recognized that chief among the potential problems was the risk that methane gas escaping from the well would come into contact with sea water and form slushy hydrates, essentially clogging the cofferdam with hydrocarbon ice.
13. BP planned to mitigate this concern once the dome had been installed by circulating warm water into the dome from the surface, so that hydrocarbons could flow up the riser unimpeded.
14. Notwithstanding these uncertainties, BP, in a presentation to the leadership of the Department of the Interior, described the probability of the cofferdam's success as Medium/High.
15. Others in the oil and gas industry were not so optimistic : Experts have told Commission staff that it was widely understood within the industry that the cofferdam effort was very likely to fail due to hydrate formation.
16. BP's effort to capture oil from the Macondo well with the containment dome did not succeed.
17. While BP had a plan to deal with hydrates once the cofferdam was in place, it had not planned to mitigate hydrate formation during the installation process itself.

18. When crews started to maneuver the cofferdam over the leak at the end of the riser on the evening of May 7, hydrates formed before the dam could be put in place, clogging the opening through which oil was to be funneled.
19. BP Vice President Richard Lynch, who oversaw the cofferdam effort, told Commission staff that BP did not anticipate hydrates forming this early.
20. Because hydrates are lighter than water, they also rendered the containment dome buoyant as it was still being lowered.
21. In the New York Times, Lynch recalled engineers telling him that they had lost the cofferdam, which, after filling with highly flammable hydrates, had begun floating up toward the ship-covered ocean surface.
22. Engineers were eventually able to gain control of the 98-ton dome and move it to safety on the sea floor.
23. One high-level government official recalled Andy Inglis, BP's Chief Executive of Exploration & Production, saying: if we had tried to make a hydrate collection contraption, we couldn't have done a better job.
24. The lack of an accurate flow-rate estimate may have hindered BP's planning for the cofferdam.
25. Suttles told Commission staff that, at the time BP deployed the cofferdam, no one at BP believed the flow was greater than 13-14000 barrels per day (bbls/day).
26. The government's then-current estimate of the flow rate was 5000 bbls/day, an order of magnitude lower than its now-current estimate of the flow in early May (approximately 60000 bbls/day).
27. Government officials have told Commission staff that part of the reason for the quicker-than-expected formation of hydrates in the cofferdam was the large flow volume.
28. Moreover, BP had publicly predicted that the cofferdam would remove about 85 % of the oil spilling into the sea.
29. But the ship BP planned to connect to the cofferdam, the Discoverer Enterprise, was capable of processing a maximum of 15000 bbls/day.
30. If even half of the government's now-estimated 60000 bbls/day was then flowing, the containment dome could not have collected 85 % of the oil from the Macondo well, putting aside the issue of hydrates.
31. It is unclear whether a more accurate sense of the cofferdam's likelihood of success would have enabled BP to proceed differently.
32. At the time, other containment options had not yet been developed, and the cofferdam did not risk damaging the well or otherwise making the spill worse.
33. Several BP executives indicated that the Discoverer Enterprise was the only collection ship available, suggesting that a better understanding of the flow volume would not have resulted in more processing capacity for the operation.
34. Nonetheless, BP modeled hydrate formation and assessed the cofferdam's collection abilities without an accurate estimate of the oil flow.
35. Government oversight of the cofferdam operation was similar to oversight of efforts to actuate the BOP stack.

36. MMS and the Coast Guard formally approved proposed procedures, after working with BP to identify operational hazards.
37. Government officials did not substantively review BP's plan to mitigate hydrate formation or evaluate BP's predictions as to the cofferdam's likelihood of success.
38. More robust oversight, addressing such strategic and scientific issues, would not begin until late May.

Annexe 6 : la coréférence, pour l'optimisation du traitement algorithmique

Nous proposons d'utiliser l'annotation de coréférence comme puissant outil de réduction et de simplification du domaine de définition et par conséquent de facilitation pour la compréhension de la question posée. La coréférence permet de relier deux unités textuelles qui désignent la même entité (Cohen *et al.*, 2017) et (van Deemter and Kibble, 1999). Une association entre un nom et un pronom utilisé dans la même phrase ou dans la phrase suivante est typiquement une coréférence. Il s'agit dans ce cas d'une coréférence grammaticale. On pourra se baser sur l'algorithme de l'université de Stanford[249] pour retrouver les coréférences éventuelles dans une phrase, mais aussi entre des phrases consécutives.

La coréférence au sein d'une phrase :

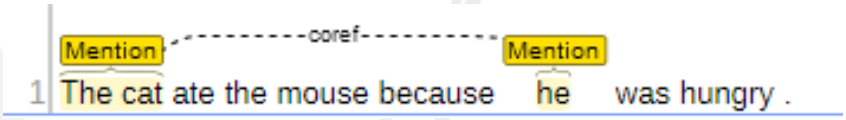


Figure 115 : coréférence entre un groupe nominal et un pronom

Le pronom *he* reprend bien le groupe nominal (syntagme nominal) *The cat*.

A partir de là, nous pensons qu'il peut être intéressant par la suite de trouver un moyen de remplacement automatique des pronoms par ce qu'ils représentent ce qui permettrait de réécrire la phrase (en tout état de cause que la machine que nous allons déployer la lise comme tel) de la sorte :

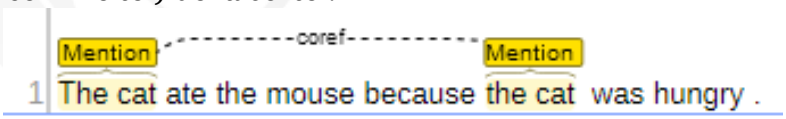


Figure 116 : le pronom est remplacé par ce qu'il représente

Ce faisant, trois résultats intéressants peuvent être attendus :

- Augmenter sensiblement la « masse sémantique » des syntagmes d'intérêt,
- Simplifie drastiquement le domaine de définition (virtuellement la catégorie grammaticale PRP, pronoms, peut être entièrement éliminée),
- Augmente nettement les co-occurrences entre deux phrases.

Par exemple, en éliminant cette coréférence grammaticale, on éclate la masse sémantique du pronom *he* (qui est indéfini) en une multitude de renforcements des masses sémantiques des syntagmes par essence déjà définis.

Nous proposons d'étendre la portée de la coréférence à des groupes de mots qui possèdent selon notre expertise du domaine une similarité forte dans leur extension sémantique avec pour objectif de faire apparaître la dénotation[250] dans le domaine de définition. L'objectif est de pouvoir notamment relier des entités entre deux phrases différentes, mais qui mentionnent « la même chose ». La méthode d'association serait :

---

249 Source : (Lee et al., 2013).

250 Définition donnée par le CNRTL : LOG. Dénotation d'un concept, d'un terme. Ce qui correspond à son extension ou l'ensemble de ses sèmes génériques. LINGUISTIQUE [P. oppos. à connotation] Ensemble des traits distinctifs qui objectivement caractérisent cette classe. ("DÉNOTATION : Définition de DÉNOTATION," n.d.).

- S'il s'agit bien de la même entité, la « *même représentation mentale de la référence dans le texte* » (Reboul and Gaiffe, 1999)
- Alors, il s'agit de la désigner strictement de la même manière dans l'ensemble du domaine de définition.

Déterminer les coréférences permet de pouvoir « *résoudre la référence* » *(Popescu-Belis, 1998 ; Popescu-Belis et al., 1998)* et d'amener le lecteur, humain ou machine à comprendre ce qu'il lit.

Une fois établi l'ensemble des coréférences d'abord grammaticales puis par extensions sémantiques identiques dans les phrases ou entre phrases, nous pouvons « réécrire le texte » en utilisant une seule et unique forme pour chaque référence et coréférences associées. Typiquement, on aurait une approche de la sorte. Soit trois phrases :

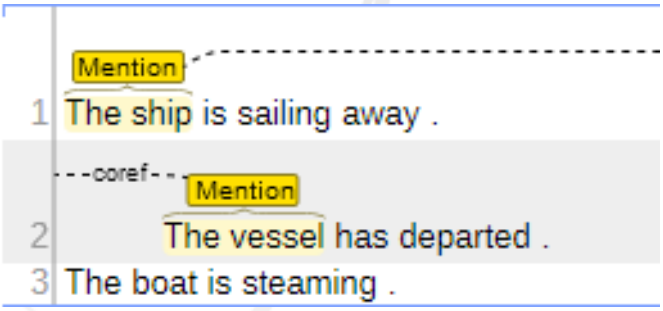


Figure 117 : coréférences inter phrases

La première question est de savoir si, dans notre domaine de définition, *The ship* et *The vessel* désignent bien la même entité. Admettons que cela soit le cas. De la même manière, posons-nous la question de savoir si *The boat* désigne la même entité que les deux autres ; si c'est le cas, il faut créer la coréférence, sinon, il n'y en a pas. Admettons que cela ne soit pas le cas.

Alors, pour faire « tomber une dimension » de l'espace sémantique associé aux deux unités, on pourra réécrire ces trois phrases de la sorte :

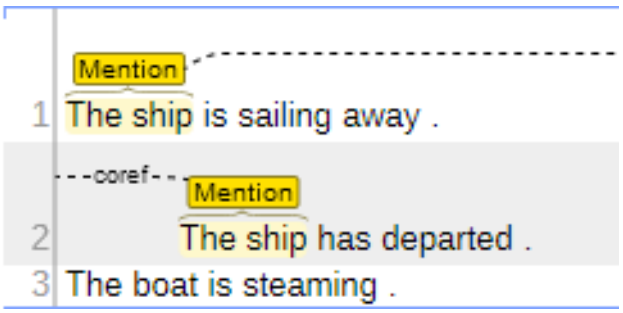


Figure 118 : réécriture par synonymie

Nous avons réécrit cette phrase en changeant une unité par une autre en utilisant le lien de synonymie. Mais il pourrait en être autrement ; voyons l'exemple suivant :

The Queen Mary 2 has left the harbour. The ocean liner will reach the next port in a few days.

Il y a deux groupes possibles *a priori* :

The Queen Mary 2 avec *The ocean liner*

et

the harbour avec *the next port*

Un humain comprend vite que la deuxième proposition n'a pas de sens malgré la synonymie possible entre *harbour* et *port*.

Quant au premier groupe, si un humain sait que *The Queen Mary 2* est un navire et particulièrement un *ocean liner,* alors il peut associer les deux parce qu'il a la même représentation mentale des deux expressions qui rappellent *in fine* la même entité, il a résolu la coréférence[251]. La suite de ce travail de résolution sera donc de choisir comment exprimer dans le texte cette entité. Annoter les coréférences apporte un degré de connaissance supplémentaire au texte :

- en créant des ensembles (des classes de « mots ») similaires ou considérés comme tel,
- en créant des « fils conducteurs » par le truchement de ces classes dans le texte étudié. On crée ainsi des chaînes sémantiques entre les phrases.

On pourra constituer un ensemble de cliques d'unités lexicales qui désignent les mêmes entités dans la limite du domaine de définition. Ce type de travail nécessite pour l'humain un degré de connaissance suffisant pour être capable de proposer des associations pertinentes entre des formes du texte et sera probablement fastidieux. L'objectif final est que les différents traitements algorithmiques que nous avons mis au point dans ce chapitre puissent voir leurs résultats sérieusement augmentés.

[251] Notons dans ce cas que l'algorithme de Stanford ne détecte aucune coréférence entre ces deux phrases.

Annexe 7 : le tableau d'annotation manuelle des phrases causales et non-causales du texte *staff working paper* n°6[252]

Il y a 553 phrases que nous avons donc classifiées en causale (**C**) ou non-causale (**N**). Le numéro est celui d'apparition de la phrase dans le texte.

phrase number	causal	phrase number	non causal	phrase number	non causal	phrase number	non causal	phrase number	non causal	phrase number	non causal
4	C	1	N	116	N	246	N	370	N	496	N
10	C	2	N	117	N	247	N	371	N	497	N
11	C	3	N	118	N	248	N	372	N	498	N
12	C	5	N	120	N	249	N	373	N	500	N
15	C	6	N	121	N	250	N	374	N	501	N
37	C	7	N	122	N	251	N	375	N	503	N
38	C	8	N	123	N	252	N	376	N	505	N
53	C	9	N	126	N	253	N	377	N	506	N
54	C	13	N	127	N	254	N	378	N	507	N
102	C	14	N	128	N	255	N	380	N	509	N
103	C	16	N	129	N	256	N	381	N	510	N
105	C	17	N	130	N	257	N	382	N	512	N
108	C	18	N	133	N	258	N	384	N	513	N
109	C	19	N	134	N	259	N	386	N	514	N
115	C	20	N	135	N	260	N	387	N	515	N
119	C	21	N	136	N	261	N	388	N	517	N
124	C	22	N	140	N	262	N	389	N	520	N
125	C	23	N	141	N	265	N	392	N	521	N
131	C	24	N	142	N	272	N	393	N	523	N
132	C	25	N	145	N	277	N	394	N	524	N
137	C	26	N	147	N	279	N	395	N	526	N
138	C	27	N	148	N	280	N	396	N	527	N
139	C	28	N	151	N	283	N	397	N	528	N
143	C	29	N	154	N	284	N	398	N	529	N
144	C	30	N	155	N	285	N	399	N	530	N
146	C	31	N	156	N	287	N	400	N	535	N
149	C	32	N	157	N	288	N	401	N	537	N
150	C	33	N	158	N	290	N	402	N	540	N
152	C	34	N	160	N	291	N	403	N	541	N
153	C	35	N	161	N	292	N	404	N	543	N
159	C	36	N	163	N	293	N	405	N	545	N
162	C	39	N	164	N	294	N	406	N	546	N
171	C	40	N	165	N	295	N	407	N	547	N
175	C	41	N	166	N	296	N	408	N	548	N
194	C	42	N	167	N	297	N	409	N	549	N
199	C	43	N	168	N	298	N	410	N	550	N
208	C	44	N	169	N	299	N	411	N	552	N
211	C	45	N	170	N	300	N	412	N	553	N
212	C	46	N	172	N	301	N	413	N		
216	C	47	N	173	N	302	N	414	N		
218	C	48	N	174	N	303	N	415	N		

[252] (National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, 2011a)

219	C	49	N	176	N	304	N	416	N		
221	C	50	N	177	N	305	N	417	N		
224	C	51	N	178	N	306	N	418	N		
225	C	52	N	179	N	307	N	420	N		
263	C	55	N	180	N	308	N	421	N		
264	C	56	N	181	N	309	N	422	N		
266	C	57	N	182	N	310	N	423	N		
267	C	58	N	183	N	311	N	425	N		
268	C	59	N	184	N	312	N	426	N		
269	C	60	N	185	N	313	N	430	N		
270	C	61	N	186	N	314	N	431	N		
271	C	62	N	187	N	315	N	433	N		
273	C	63	N	188	N	316	N	434	N		
274	C	64	N	189	N	317	N	435	N		
275	C	65	N	190	N	319	N	436	N		
276	C	66	N	191	N	321	N	438	N		
278	C	67	N	192	N	322	N	439	N		
281	C	68	N	193	N	323	N	440	N		
282	C	69	N	195	N	324	N	441	N		
286	C	70	N	196	N	325	N	443	N		
289	C	71	N	197	N	326	N	444	N		
318	C	72	N	198	N	327	N	445	N		
320	C	73	N	200	N	328	N	447	N		
337	C	74	N	201	N	329	N	448	N		
338	C	75	N	202	N	330	N	449	N		
344	C	76	N	203	N	331	N	451	N		
345	C	77	N	204	N	332	N	453	N		
347	C	78	N	205	N	333	N	454	N		
379	C	79	N	206	N	334	N	455	N		
383	C	80	N	207	N	335	N	457	N		
385	C	81	N	209	N	336	N	458	N		
390	C	82	N	210	N	339	N	459	N		
391	C	83	N	213	N	340	N	460	N		
419	C	84	N	214	N	341	N	461	N		
424	C	85	N	215	N	342	N	466	N		
427	C	86	N	217	N	343	N	468	N		
428	C	87	N	220	N	346	N	469	N		
429	C	88	N	222	N	348	N	470	N		
432	C	89	N	223	N	349	N	471	N		
437	C	90	N	226	N	350	N	473	N		
442	C	91	N	227	N	351	N	474	N		
446	C	92	N	228	N	352	N	475	N		
450	C	93	N	229	N	353	N	476	N		
452	C	94	N	230	N	354	N	477	N		
456	C	95	N	231	N	355	N	478	N		
462	C	96	N	232	N	356	N	479	N		
463	C	97	N	233	N	357	N	480	N		
464	C	98	N	234	N	358	N	481	N		
465	C	99	N	235	N	359	N	483	N		
467	C	100	N	236	N	360	N	484	N		
472	C	101	N	237	N	361	N	485	N		
482	C	104	N	238	N	362	N	487	N		

486	C	106	N	239	N	363	N	488	N		
489	C	107	N	240	N	364	N	490	N		
499	C	110	N	241	N	365	N	491	N		
502	C	111	N	242	N	366	N	492	N		
504	C	112	N	243	N	367	N	493	N		
508	C	113	N	244	N	368	N	494	N		
511	C	114	N	245	N	369	N	495	N		
516	C										
518	C										
519	C										
522	C										
525	C										
531	C										
532	C										
533	C										
534	C										
536	C										
538	C										
539	C										
542	C										
544	C										
551	C										

Annexe 8 : le fichier de classification généré par *KhCoder* pour le traitement d'un document nouveau, la phase de détection de phrases causales

1. Le document n°1 Britannica

	frequency	n	c	variance	n (%)	c (%)
the-OTHER	4	27,35	27,27	0	50,07	49,93
,-OTHER	3	19,21	18,43	0,15	51,04	48,96
in-OTHER	3	15,3	14,35	0,22	51,6	48,4
and-OTHER	2	11,04	10,48	0,08	51,32	48,68
[prior probability]	1	9,27	7,93	0,45	53,89	46,11
by-OTHER	2	7,82	6,39	0,52	55,06	44,94
oil-Noun	2	7,17	7,37	0,01	49,3	50,7
,-OTHER	1	6,04	5,55	0,06	52,11	47,89
be-Verb	1	5,46	5,38	0	50,37	49,63
a-OTHER	1	5,33	5,26	0	50,32	49,68
BP-ProperNoun	1	5,31	4,94	0,03	51,81	48,19
company-Noun	2	3,89	3,79	0	50,68	49,32
operate-Verb	1	2,56	2,4	0,01	51,62	48,38
Horizon-ProperNoun	1	2,3	1,49	0,17	60,74	39,26
Deepwater-ProperNoun	1	2,3	1,49	0,17	60,74	39,26
rig-Noun	1	2,2	0,79	0,49	73,43	26,57
Transocean-ProperNoun	1	1,79	2,4	0,09	42,7	57,3

2. Le document n°2 Slate

	frequency	n	c	variance	n (%)	c (%)
the-OTHER	4	27,35	27,27	0	50,07	49,93
be-Verb	2	10,92	10,76	0,01	50,37	49,63
[prior probability]	1	9,27	7,93	0,45	53,89	46,11
,-OTHER	1	6,4	6,14	0,02	51,04	48,96
into-OTHER	2	6,27	8,46	1,2	42,58	57,42
,-OTHER	1	6,04	5,55	0,06	52,11	47,89
of-OTHER	1	5,66	5,59	0	50,32	49,68
a-OTHER	1	5,33	5,26	0	50,32	49,68
in-OTHER	1	5,1	4,78	0,02	51,6	48,4
on-OTHER	1	4,82	4,01	0,16	54,56	45,44
which-W	1	3,53	2,87	0,11	55,09	44,91
about-OTHER	1	2,83	2,59	0,02	52,27	47,73
call-Verb	1	2,48	1,49	0,25	62,54	37,46
Horizon-ProperNoun	1	2,3	1,49	0,17	60,74	39,26
Deepwater-ProperNoun	1	2,3	1,49	0,17	60,74	39,26
drilling-Noun	1	2,2	2,4	0,01	47,75	52,25

member-Noun	1	2,08	1,89	0,01	52,34	47,66
new-Adj	1	2,08	2,59	0,06	44,56	55,44
there-OTHER	1	1,79	2,74	0,23	39,53	60,47
minute-Noun	1	1,61	0,79	0,17	66,94	33,06
same-Adj	1	1,39	1,49	0	48,23	51,77
area-Noun	1	1,39	1,49	0	48,23	51,77
name-Noun	1	1,39	0,79	0,09	63,55	36,45
40-OTHER	1	0,69	0,79	0	46,58	53,42
crew-Noun	1	0,69	1,49	0,16	31,78	68,22

Annexe 9 : l'empreinte syntaxique

La phrase est :« *BP planned to mitigate this concern once the dome had been installed by circulating warm water into the dome from the surface, so that hydrocarbons could flow up the riser unimpeded.* »

Nombre de Relation		Relation																		
FROM word	TO word	advcl	amod	aux	auxpass	case	cc	compoun	compoun	conj : so	det	dobj	mark	nmod :	nmod :	nsubj	nsubj :	nsubjpass	root	xcomp	Total
circulating-14	by-13												1								1
	dome-19														1						1
	water-16											1									1
Total circulating-14												1	1		1						3
concern-6	this-5										1										1
Total concern-6											1										1
dome-19	into-17					1															1
	surface-													1							1

	the-18										1										1
Total dome-19						1					1			1							3
dome-9	the-8										1										1
Total dome-9											1										1
flow-28	could-27			1																	1
	hydrocarb															1					1
	that-25												1								1
	unimpede											1									1
	up-29								1												1
Total flow-28				1					1			1	1			1					5
installed-12	been-11				1																1
	circulatin	1																			1

																			1			1
										1												1
				1																		1
													1									1
							1															1
		1		1	1		1			1			1					1			7
mitigate-4	BP-1																1					1
	concern-6											1										1
	flow-28	1																				1
	installed-	1																				1
	to-3												1									1
Total mitigate-4		2										1	1				1					5

water-16	Total unimpeded-32	unimpeded-32		Total surface-22	surface-22		Total ROOT-0	ROOT-0	Total planned-2	planned-2	
warm-15		the-30	riser-31		the-21	from-20		planned-2		mitigate-4	BP-1
1											
				1		1					
	1		1								
	1	1		1	1						
									1		1
							1	1			
									1	1	
1	2	1	1	2	1	1	1	1	2	1	1

Total water-16			1																			1
Total général		3	1	2	1	2	1	1	1	1	5	3	4	1	1	2	1	1	1	1		33

RÉSUMÉ

Le forage de données, méthode et moyens développés dans cette thèse, redéfinit le processus d'extraction de données, de la formalisation de la connaissance et de son enrichissement notamment dans le cadre de l'élucidation d'évènements qui n'ont pas ou peu été documentés. L'accident de la plateforme de forage *Deepwater Horizon*, opérée pour le compte de BP dans le Golfe du Mexique et victime d'un *blowout* le 20 avril 2010, sera notre étude de cas pour la mise en place de notre preuve de concept de forage de données. Cet accident est le résultat d'un décalage inédit entre l'état de l'art des heuristiques des ingénieurs de forage et celui des ingénieurs antipollution. La perte de contrôle du puits MC 252-1 est donc une faillite d'ingénierie et il faudra quatre-vingt-sept jours à l'équipe d'intervention pour reprendre le contrôle du puits devenu sauvage et stopper ainsi la pollution. *Deepwater Horizon* est en ce sens un cas d'ingénierie en situation extrême, tel que défini par Guarnieri et Travadel.

Nous proposons d'abord de revenir sur le concept général d'accident au moyen d'une analyse linguistique poussée présentant les espaces sémantiques dans lesquels se situe l'accident. Cela permet d'enrichir son « noyau de sens » et l'élargissement de l'acception commune de sa définition.

Puis, nous amenons que la revue de littérature doit être systématiquement appuyée par une assistance algorithmique pour traiter les données compte tenu du volume disponible, de l'hétérogénéité des sources et des impératifs d'exigences de qualité et de pertinence. En effet, plus de huit cent articles scientifiques mentionnant cet accident ont été publiés à ce jour et une vingtaine de rapports d'enquêtes, constituant notre matériau de recherche, ont été produits. Notre méthode montre les limites des modèles d'accidents face à un cas comme *Deepwater Horizon* et l'impérieuse nécessité de rechercher un moyen de formalisation adéquat de la connaissance.

De ce constat, l'utilisation des ontologies de haut niveau doit être encouragée. L'ontologie *DOLCE* a montré son grand intérêt dans la formalisation des connaissances à propos de cet accident et a permis notamment d'élucider très précisément une prise de décision à un moment critique de l'intervention. La population, la création d'instances, est le cœur de l'exploitation de l'ontologie et son principal intérêt mais le processus est encore très largement manuel et non exempts d'erreurs. Cette thèse propose une réponse partielle à ce problème par un algorithme *NER* original de population automatique d'une ontologie.

Enfin, l'étude des accidents n'échappe pas à la détermination des causes et à la réflexion sur les « faits socialement construits ». Cette thèse propose les plans originaux d'un « pipeline sémantique » construit à l'aide d'une série d'algorithmes qui permet d'extraire la causalité exprimée dans un document et de produire un graphe représentant ainsi le « cheminement causal » sous-jacent au document. On comprend l'intérêt pour la recherche scientifique ou industrielle de la mise en lumière ainsi créée du raisonnement afférent de l'équipe d'enquête. Pour cela, ces travaux exploitent les avancées en *Machine Learning* et *Question Answering* et en particulier les outils *Natural Language Processing*.

Cette thèse est un travail d'assembleur, d'architecte, qui amène à la fois un regard premier sur le cas *Deepwater Horizon* et propose le forage des données, une méthode et des moyens originaux pour aborder un évènement, afin de faire émerger du matériau de recherche des réponses à des questionnements qui échappaient jusqu'alors à la compréhension.

MOTS CLÉS

Deepwater Horizon, accident, ontologies, DOLCE, causalité, apprentissage automatique

ABSTRACT

Data drilling, the method and means developed in this thesis, redefines the process of data extraction, the formalization of knowledge and its enrichment, particularly in the context of the elucidation of events that have not or only slightly been documented. The Deepwater Horizon disaster, the drilling platform operated for BP in the Gulf of Mexico that suffered a blowout on April 20, 2010, will be our case study for the implementation of our proof of concept for data drilling. This accident is the result of an unprecedented discrepancy between the state of the art of drilling engineers' heuristics and that of pollution response engineers. The loss of control of the MC 252-1 well is therefore an engineering failure and it will take the response party eighty-seven days to regain control of the wild well and halt the pollution. Deepwater Horizon is in this sense a case of engineering facing extreme situation, as defined by Guarnieri and Travadel.

First, we propose to return to the overall concept of accident by means of an in-depth linguistic analysis presenting the semantic spaces in which the accident takes place. This makes it possible to enrich its "core meaning" and broaden the shared acceptance of its definition.

Then, we bring that the literature review must be systematically supported by algorithmic assistance to process the data taking into account the available volume, the heterogeneity of the sources and the requirements of quality and relevance standards. In fact, more than eight hundred scientific articles mentioning this accident have been published to date and some twenty investigation reports, constituting our research material, have been produced. Our method demonstrates the limitations of accident models when dealing with a case like Deepwater Horizon and the urgent need to look for an appropriate way to formalize knowledge.

As a result, the use of upper-level ontologies should be encouraged. The DOLCE ontology has shown its great interest in formalizing knowledge about this accident and especially in elucidating very accurately a decision-making process at a critical moment of the intervention. The population, the creation of instances, is the heart of the exploitation of ontology and its main interest, but the process is still largely manual and not without mistakes. This thesis proposes a partial answer to this problem by an original NER algorithm for the automatic population of an ontology.

Finally, the study of accidents involves determining the causes and examining "socially constructed facts". This thesis presents the original plans of a "semantic pipeline" built with a series of algorithms that extract the expressed causality in a document and produce a graph that represents the "causal path" underlying the document. It is significant for scientific or industrial research to highlight the reasoning behind the findings of the investigation team. To do this, this work leverages developments in Machine Learning and Question Answering and especially the Natural Language Processing tools.

As a conclusion, this thesis is a work of a fitter, an architect, which offers both a prime insight into the Deepwater Horizon case and proposes the data drilling, an original method and means to address an event, in order to uncover answers from the research material for questions that had previously escaped understanding.

KEYWORDS

Deepwater Horizon, accident, ontologies, DOLCE, causality, machine learning