Identifier les traces pertinentes dans la documentation scientifique d'une entreprise à des fins de communication institutionnelle. Le cas de la Compagnie TotalEnergies.

Charlotte Darricades, Sébastien Laborie, Christian Sallaberry, Eric Kergosien, Patrice de La Broise

Identifier les traces pertinentes dans la documentation scientifique d'une entreprise à des fins de communication institutionnelle.

Le cas de la Compagnie TotalEnergies

Charlotte Darricades[1,3], Christian Sallaberry[2] , Sébastien Laborie[2], Eric Kergosien[1], Patrice De La Broise[1]

1. Universite de Lille, GERiiCO
charlotte.darricades@univ-lille.fr, eric.kergosien@univ-lille.fr, patrice.de-la-broise@univ-lille.fr
2. Universite de Pau et des Pays de l'Adour, E2S UPPA, LIUPPA
christian.sallaberry@univ-pau.fr, sebastien.laborie@univ-pau.fr
3. Pôle d'Études et de Recherche de Lacq, TotalEnergies SE, BP 47, 64170 Lacq
charlotte.darricades@totalenergies.com

Résumé: Nous faisons l'hypothèse que des traces d'activités thématiques présentes dans les productions scientifiques d'une entreprise (le cas de TotalEnergies) sont utiles à des fins de développement social d'une entreprise en pleine mutation. Nous présentons une nouvelle approche permettant d'identifier ces traces par l'exploitation d'une cartographie qui fait état de l'organisation de l'entreprise, la mise en œuvre d'entretiens avec les chercheurs et les communicants. Cela nous permettra ensuite de construire un premier squelette d'ontologie comme le préconise la méthodologie SAMOD. Enfin, nous détecterons et analyserons des traces dans les corpus documentaires scientifiques.

Mots-clés : Corpus textuels, traces d'activité scientifique, communication institutionnelle

1. Introduction

Afin de gagner l'adhésion de ses collaborateurs et du grand public face à ses mutations, chaque entreprise doit bâtir une stratégie de communication. Mettre la R&D au cœur de cette stratégie pourrait permettre d'atteindre efficacement cet objectif et permettrait d'ancrer les projets de R&D dans de nouveaux processus de communication. Les grandes entreprises qui disposent de services R&D engendrent une grande variété de documents scientifiques. C'est le cas par exemple pour la Compagnie TotalEnergies où différents départements de la R&D produisent des publications scientifiques, des rapports d'études, des brevets, etc. En parallèle à cette activité, les communicants doivent mettre en valeur l'activité de Recherche dans la stratégie de communication institutionnelle. En effet, dans une démarche de valorisation de la R&D d'une entreprise, le service Communication du Pôle d'Études et de Recherche de Lacq (PERL) de la Compagnie doit pouvoir faire émerger des données pertinentes de façon automatique, comme par exemple des thématiques en lien avec la transition énergétique telles que la Capture et le Stockage du Carbone (CCS), Biogaz, Agrivoltaisme, etc.

D'un côté, *l'information scientifique et technique (IST) est indispensable au travail des chercheurs, plus particulièrement à la construction de leur communication scientifique* (Gardiès et Fabre, 2009). D'un autre côté, elle est également indispensable pour la stratégie de communication générale d'une entreprise. Les traces dans les productions scientifiques quelles qu'elles soient peuvent par conséquent être utiles à la rédaction d'articles de stratégie de communication. Se pose alors les problématiques suivantes : Quelles sont les différentes traces pertinentes liées aux productions scientifiques ? Comment les communicants peuvent-ils exploiter ces traces d'informations de R&D qui leurs seront utiles à des fins de stratégie de communication institutionnelle ?

Dans cet article, nous proposons d'expérimenter une nouvelle approche globale permettant d'identifier des traces : de la prise en compte de l'organisation de l'entreprise jusqu'à la production d'indicateurs pour les communicants. Cette approche débute par l'exploitation d'une cartographie qui fait état de l'organisation de l'entreprise, des flux et des documents scientifiques existants. Une méthode de travail est ensuite déclinée pour permettre l'analyse du corpus documentaire.

Dans la section 2, nous allons tout d'abord aborder les méthodologies de construction d'une base de connaissance. Puis, dans la section 3, nous proposons une approche qui se décline en 2 phases distinctes, un état des lieux sous forme de cartographie suivi d'entretiens avec les chercheurs et les communicants. Enfin, nous conclurons en présentant quelques perspectives dans la section 4.

2. Travaux connexes : méthodologies pour la construction d'une base de connaissance

Nous cherchons à créer une ontologie métier pour représenter les connaissances contenues dans les différents documents scientifiques produits par les équipes du

PERL et que différents types d'utilisateurs seraient susceptibles de rechercher (communicants et chercheurs du PERL principalement). Nous avons ainsi besoin d'une méthodologie de création d'une ontologie métier extensible, qui prenne en compte les besoins des utilisateurs experts.

Un nombre important de travaux propose une méthodologie pour construire une ontologie. Parmi ceux-ci, nous pouvons notamment citer les méthodes Tove (Grüninger et Fox, 1995), Methontology (Fernandez-Lopez et al., 1997), Sensus (Swartout et al., 1997), Otk (Staab and al., 2001), Terminae (Aussenac-Gilles et al., 2008), NeOn (Suàrez-Figueroa et al., 2012), ou encore Samod (Peroni, 2016). Toutes ces méthodes commencent par une phase d'acquisition de connaissances du domaine ou du métier, de rédaction de spécifications fonctionnelles ou de questions de compétence. Methontology et Otk sont très similaires. Les deux commencent par l'acquisition de connaissances et la rédaction de spécifications. Elles se poursuivent en modélisant le domaine d'abord d'une manière informelle puis dans un langage formel. Enfin, les deux proposent une évaluation de l'ontologie produite. Methontology recommande d'ailleurs un guide d'évaluation publié dans un document annexe. NeOn et Samod proposent des consignes pour créer des ontologies modulaires. Certaines méthodologies, notamment Otk, NeOn et Samod proposent un développement modulaire de l'ontologie qui est construite petit à petit, soit en ajoutant, à chaque itération, la modélisation d'une partie supplémentaire du métier/domaine, soit en modélisant toutes les parties du métier/domaine d'abord, puis en les fusionnant. NeOn se distingue des autres méthodologies présentées car elle fournit de nombreuses approches pour élaborer une ontologie ou un réseau ontologique. Elle demande aux ontologues de réaliser préalablement une analyse approfondie du projet afin de pouvoir choisir la bonne combinaison des processus et activités proposés. Les trois méthodes Otk, NeOn et Samod intègrent également une phase d'évaluation de l'ontologie produite, et cela lors de l'étape finale. Samod semble ressortir du lot selon nos critères car elle propose une première étape permettant de créer un premier squelette d'ontologie à partir des besoins exprimés par les utilisateurs cibles. La méthode préconise d'impliquer fortement les experts concernés afin de préciser et d'étendre les besoins, et le modèle ontologie produit, de façon itérative. L'aspect itératif impliquant les experts est primordial dans un secteur spécifique tel que le nôtre. Enfin, Samod intègre des phases de tests à différentes étapes du processus. À termes, nous prévoyons ainsi de formaliser des requêtes informelles exprimées par les experts en requêtes SPARQL afin de tester à la fois le modèle ontologique, et celui-ci une fois peuplé par les données collectées dans les différentes sources de données. Nous sommes encore dans une phase d'analyses et de tests des différentes méthodes existantes, et nous confirmerons notre choix une fois ce travail d'analyse terminé.

3. Contribution : Construction d'une méthodologie test issue de différentes méthodes existantes

Dans un premier temps, nous allons produire un état des lieux à travers un sociogramme (cartographie) qui fait état de l'organisation actuelle du PERL et de sa

structure documentaire. Il permet d'en faire ressortir ses productions scientifiques avec un corpus de documents hétérogènes. Puis, nous allons identifier les besoins des utilisateurs, soit des chercheurs et des communicants des différents départements de la R&D en organisant des entretiens avec eux. Cette méthode nous permettra depuis une cartographie de produire un plan et une base de données afin d'offrir un outil utile aux communicants. Ces premiers travaux nous permettront ensuite de construire un premier squelette d'ontologie comme le préconise la méthodologie SAMOD. La particularité ici est que nous travaillons à l'enrichissement d'une cartographie au fur et à mesure des entretiens menés, afin d'identifier et formaliser précisément les concepts à modéliser, les données du corpus à mobiliser pour instancier l'ontologie, et les cas d'usages pour tester la robustesse de la base de connaissance produite.La première étape de la méthode SAMOD consistant à collecter et formaliser les besoins des experts sont présentées sections 3.1 et 3.2. L'étape 2 de la méthode est détaillée section 3.3 et la dernière étape de tests est présentée section 3.4.

3.1. Etat des lieux

La figure 1 (Annexe 1) présente donc un extrait de cette cartographie. Ainsi, le PERL dépend de plusieurs directions dans l'organigramme de TotalEnergies. De nombreuses informations au sujet de la R&D circulent entre les directions. Chaque direction (Direction OT, R&D Lines, UP Line, PERL) possède un service de communication représenté ici par 2 emojis bleu et rouge.

Le PERL a une communauté de chercheurs considérable (environ 80 chercheurs) dans plusieurs domaines. Beaucoup de documents scientifiques circulent tels que le rapport annuel, les rapports de projets de R&D, les brevets et les demandes d'inventions, les notes de synthèse R&D, et les publications scientifiques.

Prenons l'exemple des publications scientifiques. Nous allons donc nous intéresser à cette partie de la cartographie afin d'observer les types de publications scientifiques existants rédigés par les chercheurs PERL. Premièrement, il y a les articles dits « normaux », on peut dire que ces articles sont les plus communs dans le domaine de la communication scientifique. Ils exposent les résultats obtenus à l'issue d'une proposition de méthodologie et de son expérimentation. Deuxièmement, il y a les « articles de revue » où le chercheur établit un état de l'art relatif à sa problématique afin de donner une cohérence à son sujet. Troisièmement, il y les « articles de perspectives » où le chercheur-auteur va établir une revue en y ajoutant des pistes de solutions. Pour finir, les articles de "correction ou ajout" qui viennent compléter un article dit "normal" précédemment publié. À savoir que ces publications sont généralement rédigées en anglais ou en français, d'où les drapeaux bleu et rouge sur cette cartographie. Ainsi, le corpus de publications scientifiques est varié, il présente également une forte hétérogénéité de structures. De plus, il est chargé en traces thématiques de R&D qui pourraient potentiellement intéresser les communicants. Il est important de récolter toutes les traces afin de ne pas faire d'impasse et d'obtenir le plus de données sur la R&D.

3.2. Construction d'un premier modèle ontologique à partir d'entretiens

Afin de construire un premier modèle ontologique, nous organisons des entretiens semi-directifs en face à face avec des chercheurs de chaque service du PERL ainsi que des communicants de chaque direction précédemment évoquée. Nous allons donc préalablement préparer une grille d'entretiens et mettre en forme la première cartographie pour être le plus précis possible durant l'échange avec chaque expert. Ces entretiens auront tout d'abord pour but d'aligner le vocabulaire scientifique au vocabulaire des communicants. Ensuite, ils nous permettront de relever l'ensemble des traces qui sont utiles selon le chercheur et selon le communicant pour la stratégie de communication institutionnelle, et enfin de formaliser de façon claire les attentes des communicants en termes de vulgarisation de la communication scientifique. Ces entretiens semi-directifs nous offrent ainsi l'opportunité de connaître le rôle des chercheurs et l'influence de leurs publications dans la stratégie de communication institutionnelle. À partir de l'ensemble de ces éléments, nous allons pouvoir construire un premier squelette de l'ontologie, que nous pourrons ensuite enrichir et instancier.

3.3. Extraction d'informations pour l'instanciation de l'ontologie

Comme nous l'avons décrit précédemment, nous disposons d'un corpus documentaire hétérogène. Également, nous disposons d'une cartographie qui permet de connaître les producteurs de ces documents ainsi que leurs relations entre eux (ex. relations hiérarchiques, relations de collaborations…). Enfin, le communiquant désireux d'analyser le corpus dispose d'indicateurs qui vont lui permettre d'ajuster sa politique de communication. Une première stratégie pour produire ces indicateurs consisterait à indexer en amont tout le corpus documentaire sans tenir compte d'un contexte d'analyse donné. Il va de soi que cette méthode n'est pas efficace car d'une part elle pourrait produire des indicateurs non-utiles au communiquant à un instant t, et d'autre part certains indicateurs pourraient ne pas être pertinents lorsqu'ils sont appliqués de façon globale sur certains types de documents. Par conséquent, tenant compte des différents types de documents ainsi que de notre cartographie, nous proposons plutôt une seconde stratégie qui consiste à cibler l'indexation en fonction d'un contexte d'analyse du communiquant. Par exemple, si le communiquant désire avoir une vue globale de son corpus au sujet des tendances de thématiques abordées au sein de son organisation, un indexeur spécifique pourra être sélectionné plutôt que tous les indexeurs possibles (ex., un extracteur de concepts) et certaines parties de documents pourront être analysées plus particulièrement par cet indexeur (ex., les titres et les résumés des articles scientifiques, les références bibliographiques pour les rapports d'activités…).

En effet, il existe actuellement de nombreux outils d'indexation de documents, notamment textuels, allant d'outils standards "clé en main" jusqu'à des outils configurables par des experts informaticiens. En voici quelques exemples :

- Les outils standards “clé en main” : Voyant[1], VOSViewer[2], Bibliometrics[3], Sketchengine, [4]Cortext[5], Gargantext[6]… Il s’agit globalement d’outils dédiés aux spécialistes de la langue qui désirent, via une interface graphique adaptée, pratiquer l'extraction terminologique, l'alignement multilingue, la visualisation des occurrences de termes en contexte ou encore l'édition de ressources linguistiques,
- Les outils avancés “paramétrables” : TextRazor[7], Lexalytics[8] … Ils combinent des techniques de traitement du langage naturel avec des bases de connaissances pour extraire des entités informationnelles dans des documents, tweets ou pages web. Ils peuvent être utilisés en version standard ou avec un minimum de paramétrage, tout comme certains proposent des librairies de fonctions intégrables dans des programmes ad-hoc. En version standard, ils reconnaissent, par exemple, des organisations, des dates, des lieux, des prix, des adresses, des personnages célèbres, des œuvres d’art, etc.,
- Les outils experts “avec programmation” : Gate[9], Spacy[10], Google NLP[11] … Il s’agit de boîtes à outils logicielles utilisées pour le traitement du langage naturel dans différentes langues. Différents services sont mis à disposition des programmeurs d’applications dans des langages tels que Java et Python.

Voici un exemple d’article de recherche (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4286317) qui, après traitement par des outils existants, présente déjà des résultats d’extraction intéressants en ciblant leurs traitements uniquement sur le résumé de l’article : voir Figures n°2 (Annexe 2) et n°3 (Annexe 3).

En fait, nous avons pour objectif d’associer des outils d’annotation à notre base de connaissance métier relative à l’énergie et à l’environnement. Nous devrons nous appuyer sur des outils suffisamment ouverts qui intègrent de tels types de ressources externes. Dans le cadre d’un premier travail de veille, nous avons identifié TextRazor et Gate qui proposent ce type de service. Le travail d’annotation devra nous aider à peupler semi-automatiquement l’ontologie produite en amont.

---

[1] (https://voyant-tools.org/)
[2] https://www.vosviewer.com/
[3] https://www.bibliometrix.org
[4] https://www.sketchengine.eu/
[5] https://www.cortext.net
[6] https://www.gargantext.org
[7] https://www.textrazor.com/
[8] https://www.lexalytics.com/
[9] https://gate.ac.uk/
[10] https://spacy.io/
[11] https://cloud.google.com/natural-language

3.4. Outils d'analyses et de visualisation

Dans le cadre de notre projet, chaque communicant devrait pouvoir construire son propre tableau de bord, afin d'avoir une vision globale du corpus documentaire lui facilitant, par la suite, sa prise de décision. Pour ce faire, nous pourrions :

- exploiter les fonctionnalités complémentaires des outils d'extraction identifiés plus haut comme, par exemple, Voyant qui permet l'affichage de nuage de mots et/ou VOSViewer qui présente des réseaux bibliométriques,
- ou bien, sur la base d'extractions réalisées en amont, utiliser des outils d'analyse et de visualisation externes comme, par exemple, les outils de la Business Intelligence tels que Tableau, Qlik ou Power BI. Notons l'émergence d'outils No-Code spécialisés dans le NLP : SimpleX (https://sx.simpledecisions.io/landing/about-us), par exemple, est une console de text mining dédiée au traitement et à la visualisation de données textuelles.

4. Conclusion / perspectives

Cette première méthode permettant d'identifier ces traces par l'exploitation d'une cartographie, avec la mise en œuvre d'entretiens pour construire un modèle ontologique, suivis d'une étape d'analyse des corpus documentaires scientifiques pour instancier ce modèle et obtenir une base de connaissance métier. Nous sommes actuellement dans la phase d'exploration à travers le processus de construction des connaissances avec en aval une réflexion sur la définition de l'objet de recherche (via la problématique des communicants et l'identification des sources pertinentes pour produire des supports de communication), et en amont les données (recueil des traces sur la R&D et traitement) ainsi que sur les choix finaux concernant le dispositif méthodologique (Charreire-Petit et Durieux, 2014).

Selon la méthodologie finale choisie, celle-ci pourrait permettre de saisir une nouvelle approche de la communication scientifique inter-métiers.

Bibliographie non numérotée

Aussenac-Gilles N., Despres S., Szulman S. (2008). *The terminae method and platform for ontology engineering from texts*. Paul Buitelaar and Philipp Cimiano, editors, Bridging the Gap between Text and Knowledge, pp 199-223

Charreire-Petit S., Durieux F. (2014). *Méthodes de recherche en management*. Dunod, chap. 3, pp 76-104

Fernandez-Lopez M., Gomez-Perez A., Juristo N. (1997). *METHONTOLOGY: From Ontological Art Towards Ontological Engineering* https://oa.upm.es/5484/?trk=public_post_main-feed-card-text

Gardiès C., Fabre I. (2009). *Communication scientifique et traitement documentaire de l'IST. Quelles méthodes du travail intellectuel ?* Les Cahiers du numérique, vol. 5, no. 2, pp. 85-104.

Gruninger M., Fox M. (1995). *Methodology for the Design and Evaluation of Ontologies* Workshop on Basic Ontological Issues in Knowledge Sharing

Peroni S. (2016). *SAMOD: an agile methodology for the development of ontologies* https://essepuntato.it/samod/

Skuce D. (1993). *D.B. Lenat and R.V. Guha, Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project** Artificial Intelligence Volume 61, Issue 1, pp 81-94

Staab S., Schnurr H.-P., Studer R., Sure Y. (2001). *Knowledge processes and ontologies* https://ieeexplore.ieee.org/abstract/document/912382

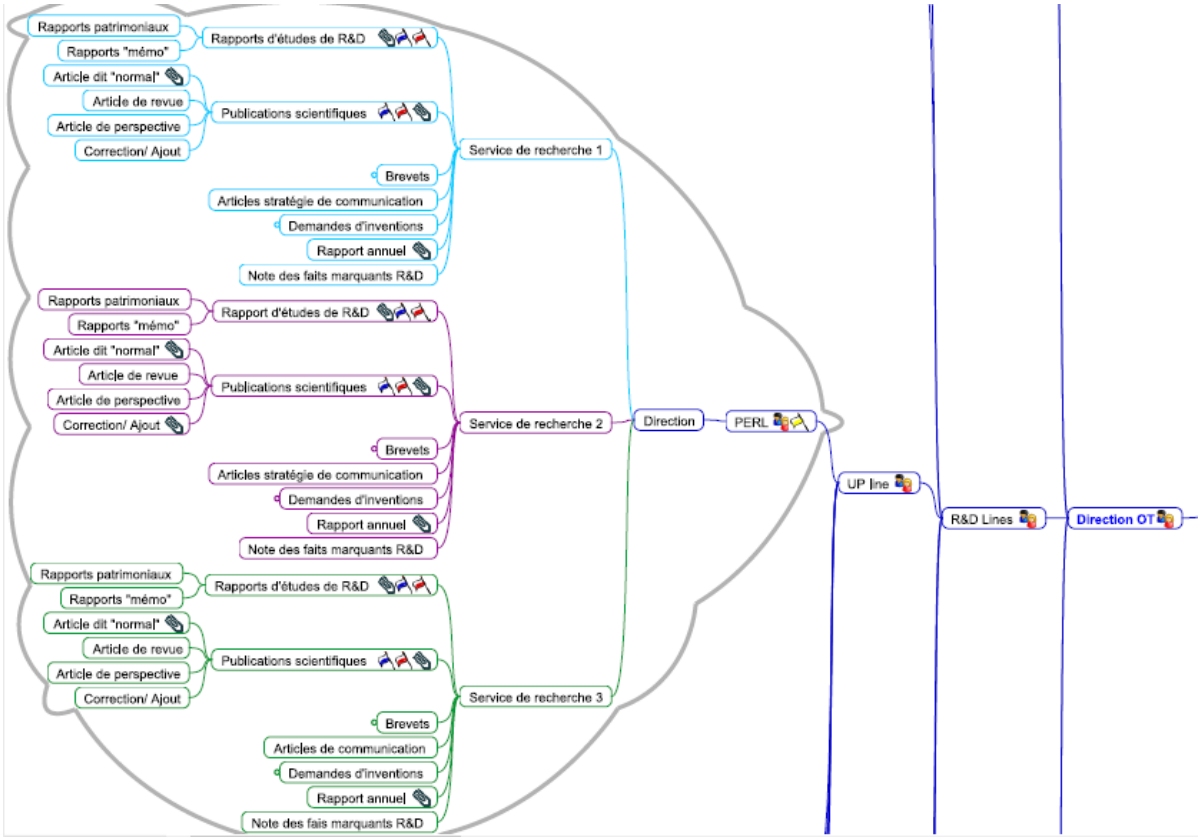
Suàrez-Figuero M.C., Gòmez-Perez A., Motta E., Gangemi A. (2012). *The NeOn Methodology for Ontology Engineering*, Springer, pp 9-34

Swartout B., Ramesh P., Knight K., Russ T. (1997). *Toward Distributed Use of Large-Scale Ontologies*. https://aaai.org/papers/0018-toward-distributed-use-of-large-scale-ontologies

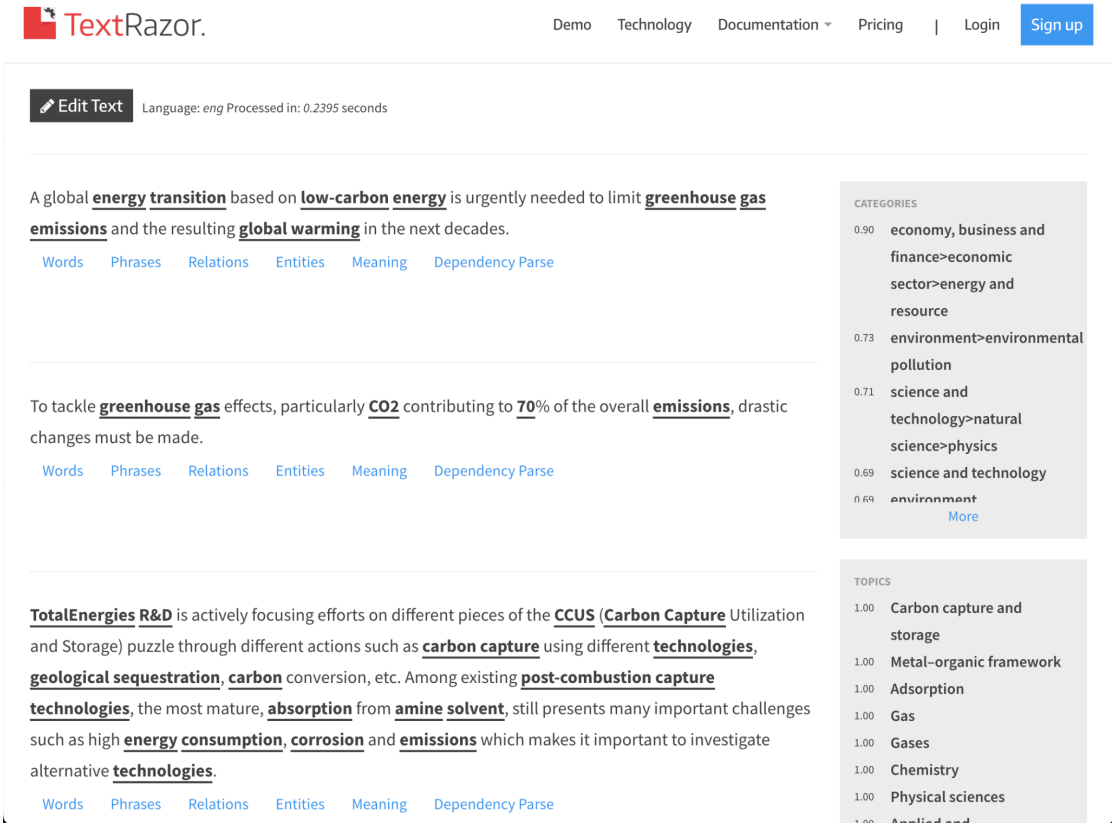
Ushold M., King M. (1995). *Towards a methodology for building ontologies* Workshop on Basic Ontological Issues in Knowledge Sharing, pp 275-280

Annexe(s)

Annexe 1. *Figure 1 : Extrait de la cartographie du PERL*



Annexe 2. *Figure 2 : Test de l'outil TextRazor*



Annexe 3. Figure 3 : *Test de l'outil Voyant*