Transformation de lexical soft data en smart data pour la construction d'un thésaurus du patrimoine minier

Amélie Daloz

Transformation de *lexical soft data* en *smart data* pour la construction d'un thésaurus du patrimoine minier

Amélie Daloz

Doctorante en Sciences de l'information et de la communication

Université de Lille, Laboratoire GERiiCO

amelie.daloz@univ-lille.fr

Résumé

L'article s'inscrit dans le cadre du projet ANR Mémo-Mines et présente le processus qui permet de transformer des *lexical soft data* en *smart data* pour la construction d'un Système d'Organisation des Connaissances (SOC). L'ouverture des données produites et structurées lors de cette transformation est un enjeu majeur. Résultat d'une méthode qualitative, l'ouverture doit permettre de sauvegarder et de valoriser un patrimoine culturel spécifique, le patrimoine minier des Hauts de France.

Mots clés

Smart data, *soft data*, organisation des connaissances, patrimoine minier, ouverture des données

Title

Transformation of lexical soft data into smart data for the building of a mining heritage thesaurus

Abstract

This article is part of the ANR Mémo-Mines project and deals with the process of transforming lexical soft data into smart data for the construction of a Knowledge Organisation System (SOC). Opening up the data produced and structured during this transformation is a major challenge. Data opening, as a result of a qualitative approach, must lead to safeguard and enhance a specific cultural heritage, the mining heritage of the Hauts de France.

Keywords

Smart data, soft data, knowledge organization, mining heritage, data opening

1 – Contexte de l'étude : le projet Mémo-Mines pour la valorisation de la mémoire minière

L'étude se situe dans le cadre du projet ANR Mémo-Mines [1], projet interdisciplinaire visant la sauvegarde des mémoires individuelles par leur conversion en traces mémorielles et leur mise à disposition sous forme de corpus d'archives numériques accessibles à tous. Les quatre champs mobilisés pour traiter la problématique sont les Sciences de l'Information et de la Communication, domaine dans lequel s'inscrivent nos travaux, les Sciences du Langage, la muséologie et l'informatique.

Si l'exploitation minière est aujourd'hui totalement arrêtée en France depuis 1990 (dernière gaillette (morceau de charbon) remontée à la fosse 9-9bis à Oignies), la volonté de sauvegarde et d'accès au patrimoine est bien présente sur le territoire et davantage depuis l'inscription sur la liste du patrimoine mondial de l'UNESCO en 2012 du Bassin Minier du Nord — Pas de Calais. Toutefois, cela implique une certaine gestion des données (de l'information et des connaissances) produites pendant et après l'exploitation. Si un certain nombre d'institutions (musées, bibliothèques et archives) se chargent du travail de médiation (documentaire, culturelle…) [2], les données produites et mises en accès libre par d'autres acteurs [3] sur Internet (blogs, pages Facebook, sites Internet, forums, Wikipédia etc.) demeurent hétérogènes de par leurs différents formats. L'un des enjeux de la recherche en cours est donc d'homogénéiser les formats et de structurer les données pour permettre un meilleur accès.

Dans ce cadre, nous définissons tout d'abord les différents types de données (dont les *lexical soft data*) dans le champ des sciences humaines et sociales et de la recherche (partie 2), puis nous exposons l'enjeu des *smart data* dans le champ de l'organisation des connaissances et du Web. En partie 3, nous appliquons les définitions aux données de notre étude et nous présentons la méthodologie de passage des *lexical soft data* aux *smart data*. En partie 3.3.b et en Annexe 1, nous proposons en résultat un échantillon du thésaurus au format d'échange adéquat grâce à un logiciel *open source* qui permet de gérer et d'exploiter des vocabulaires contrôlés, thésaurus ou taxonomies : *Tematres* [4].

2 – Data, soft data, lexical soft data et smart data en SHS

2.1 Les données *vs* les données de la recherche

Les données sont définies au niveau de base comme une absence d'uniformité (Schöch, 2013, p. 3). Dans le domaine numérique, elles sont représentées par une série de 0 et de 1. À un niveau supérieur, les données peuvent être soit linéaires (tableaux et matrices), soit hiérarchiques (avec une structure arborescente dans laquelle les éléments ont des relations parents-enfants ou frères et sœurs les uns avec les autres, comme dans un fichier XML), soit multi-relationnelles (chaque élément de données est un nœud dans un réseau interconnecté de nœuds) (Mehta et Sahni, 2005). Les données sont également caractérisées par leur niveau de structuration. Elles peuvent en effet être structurées (comme dans une base de données), semi-structurées (comme dans un fichier XML) ou non-structurées (comme dans un texte brut) (Schöch, 2013, p. 3).

Les données de la recherche sont quant à elles définies par l'OCDE (2007) comme « *des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnues par la communauté scientifique comme nécessaires pour valider des résultats de recherche.* » Le terme *data* a

largement été critiqué par les humanistes, et plusieurs définitions ont été proposées. Nous nous appuyons sur la redéfinition de Schöch (2013, p. 4) qui considère la donnée comme « *une abstraction numérique, construite de façon sélective, pouvant être actionnée par une machine et représentant certains aspects d'un objet donné de l'enquête humaniste.* » Pour l'illustrer, l'auteur divise les données en SHS en deux selon leur mode de « *création, de captation, de modélisation, d'enrichissement et d'analyse* ». Il définit ainsi les *big* et *smart data*. Avant de revenir à ce type de données, nous ajouterons une sous-division dans la catégorisation ainsi faite en parlant des *soft data* et plus précisément des *lexical soft data*. Ces dernières constituent l'objet de cette étude.

2.2 Les *soft data* et *lexical soft data* : nouveaux objets d'étude pour la recherche en SHS

Même si Terras *et al.* (2013, p. 180) ne les appellent pas *soft data*, ils distinguent le « *matériel nativement numérique » opposé aux textes numérisés (ou des sites historiques numérisés) et mentionnent le besoin d'outils « bien conçus » pour étudier « la vie en ligne et la culture* » [5].

Les données disponibles sur Internet semblent de prime abord peu exploitables ; elles sont souvent peu ou mal structurées, sont hétérogènes et sont enregistrées selon des formats divers. Elles sont néanmoins intéressantes car elles peuvent être collectées facilement. Ce type de données fait référence aux *soft data* définies par Severo et Romele (2015). Les auteurs les définissent comme « *des données disponibles sur Internet, facilement accessibles et récoltables* ». Ils ajoutent qu'« *elles sont constituées principalement par les nouveaux types de données issues du Web 2.0 (Facebook, Twitter, fils RSS, etc.) qui s'offrent au décideur public comme une source originale et riche d'informations sur les phénomènes sociaux qui ont lieu dans un territoire* ». Dans notre cas, l'approche *soft* vers *smart* permet d'analyser des données « *qui ne sont pas [forcément] libres de droit* » (Severo et Romele, 2015) mais qui sont surtout porteuses de riches informations sur les méthodes de transmission du patrimoine culturel immatériel, sur les mémoires et les souvenirs véhiculés. Dans ce contexte, nous introduisons le néologisme *lexical soft data* qui serait un type de *soft data* pouvant être doublement analysé : du point de vue de la lexicographie d'une part et du point de vue de la terminologie d'autre part en vue d'organiser les connaissances d'un domaine.

2.3 Les *smart data vs big data*

Schöch (2013, p.4) oppose les *smart data* au *big data* par leur différence de degré de structuration, leur caractère implicite ou non, leur volume et leur hétérogénéité. Ainsi, les premières seront plutôt semi-structurées ou structurées, nettoyées et explicites, en petite quantité et peu hétérogènes. Au contraire, les secondes seront relativement non structurées, désordonnées et implicites, de volume relativement important et de forme variée. Mais, comme nous l'avons vu, la véritable différence entre les deux est bien comment ces données sont créées, capturées, modelées, enrichies et analysées. Les *smart data* ont en effet tendance à ne pas être caractérisées par leur gros volume car leur création implique une « action humaine » et « demande du temps ». Nous verrons dans cette étude quel type d'action humaine est requise pour passer des *soft data* en *smart data* grâce aux *lexical soft data*.

Enfin, même si les données sont structurées, cela ne garantit pas leur interopérabilité. Dans le cadre des normes élaborées par le W3C (*World Wild Web consortium*), porté par l'inventeur du Web, Tim Berners Lee, le format SKOS [6] vise à faire migrer les thésaurus, qui sont des langages documentaires, vers des ressources disponibles sur le web sémantique. Pour cela, le format se base sur un langage de représentation des connaissances, OWL (*Web Ontology Language*), qui est lui-même construit sur le modèle de données de RDF (*Resource*

Description Framework). L'encodage de ces informations en RDF permet de faire passer les différents thésaurus d'une application informatique à l'autre d'une manière interopérable. Nous présenterons dans la partie 3.3.b un échantillon du thésaurus en format SKOS.

3 – INDEXATION ET ACCÈS AUX DONNÉES DU PATRIMOINE CULTUREL IMMATÉRIEL MINIER : DES *LEXICAL SOFT DATA* AUX *SMART DATA*

En partant du postulat que la langue est vecteur du patrimoine culturel immatériel (UNESCO, 2003) et que le numérique favorise la diffusion de ce même patrimoine par l'ouverture et le partage des données, la création d'un langage intermédiaire au langage de l'homme et de la machine est un enjeu central dans notre recherche. Dans cette optique, une double méthodologie propre à la construction d'un thésaurus pour l'indexation des ressources du domaine a été mise en place. La première (partie 3.1) consiste à construire un corpus lexical à partir de *lexical soft data* présentes sur Internet et à procéder à leur analyse terminologique et la deuxième (partie 3.2) consiste à analyser qualitativement le résultat de la première pour mettre en correspondance les différents concepts mobilisés dans le domaine minier.

3.1 Des *soft data*…

Les données du domaine minier identifiées lors d'une veille de type « pull » sur le Web sont multiples et hétérogènes (*cf.* Figure 1), il est de prime abord difficile de les capter dans leur exhaustivité. Elles peuvent en effet correspondre aussi bien à des images numérisées et modifiées qu'à des photographies numériques, à des données textuelles de tous les types (articles de presse, rapports, poèmes, paroles de chansons, notices ou références bibliographiques, commentaires, résumés, articles de blogs…), des numérisations ou des matériaux initialement numériques (cartes géographiques ou postales, sons, vidéos témoignages ou montages textes-photographies, pièces de théâtre, films, photographies, dessins, peintures, timbres,…), des noms de lieux, d'acteurs, d'événements... et proviennent de fond d'archives physiques ou virtuels, de fonds personnels ou institutionnels et sont infiniment partagées et disséminées ici et là sur la toile. Toutes ces données transportent souvent avec elles du sens textuellement explicité (légende, commentaires, notices…) mais comment capter celui-ci et surtout comment l'homogénéiser pour permettre sa valorisation ?

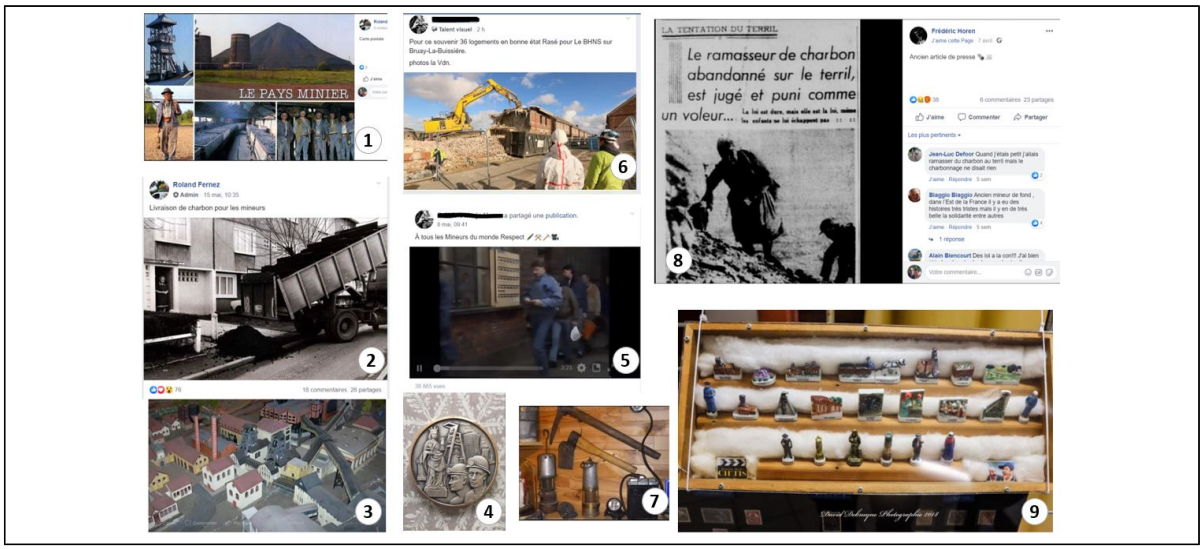

Figure 1 : Illustration de la diversité et de la richesse des *soft data* : objets de patrimoine de la page Facebook "les anciens mineurs" [(1)carte postale, (2)photographie (6)représentation d'un événement, (8)article de presse, (5)vidéo-montage, (9)fèves, (7)outils du mineur, (4)cadre, (3)maquette…]

Pour ce faire, nous avons décidé de nous intéresser aux données lexicales donc aux *lexical soft data*. La méthodologie consiste à capter celles-ci sous forme d'un corpus, de les analyser linguistiquement et de les formaliser sous forme d'une terminologie respectant les normes en vigueur, et notamment le format *TermBase eXchange* (TBX).

Grâce à l'utilisation de moteurs de recherche et à des requêtes précises, 22 ressources lexicales ont ainsi pu être identifiées provenant d'acteurs pouvant être classés ainsi :

- experts en terminologie (spécialistes de la description linguistique de la/des langues étudiée(s)),

- anciens mineurs (experts de la pratique des langues étudiées),

- passionnés (spécialistes dans le domaine étudié par leur connaissance sur le domaine).

Ce sont des lexiques, des glossaires, des dictionnaires ou des « sacs de mots ». Les *lexical soft data* ne proviennent pas de la même source (*cf.* Figure 2), n'ont pas forcément la même langue ni le même format, les termes sont plus ou moins techniques et les relations entre les différents termes ne sont pas clairement définies. L'information contenue dans ces données, pourtant riche, n'est donc pas explicitée. La diversité des acteurs et des sources (*cf.* Figures 2 et 3) est importante car elle permet ensuite de mettre en relation un maximum de termes du domaine grâce à leur analyse rendue plus automatisable par leur formalisation. Notre étude s'appuie notamment sur la méthodologie du projet TECTONIQ [7] qui étudie les dispositifs numériques mis en place par les différents acteurs impliqués pour gérer, diffuser et échanger les informations relatives au Patrimoine Industriel Textile (PIT) sur le territoire du Nord — Pas de Calais. Elle diffère néanmoins par l'utilisation d'outils très abordables par des chercheurs en SHS tels que la manipulation copier/coller pour récupérer les données et le tableur avec ses nombreuses fonctionnalités (table dynamique, fonctions, expressions régulières *etc.*) pour l'exploitation qualitative des données (*cf.* Figure 4).

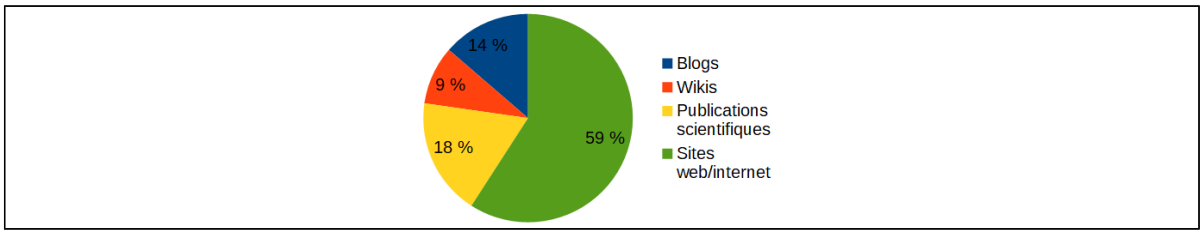

Figure 2 : Répartition quantitative et typologie des sources contenant des *lexical soft data* sur la mine du Nord —Pas de Calais

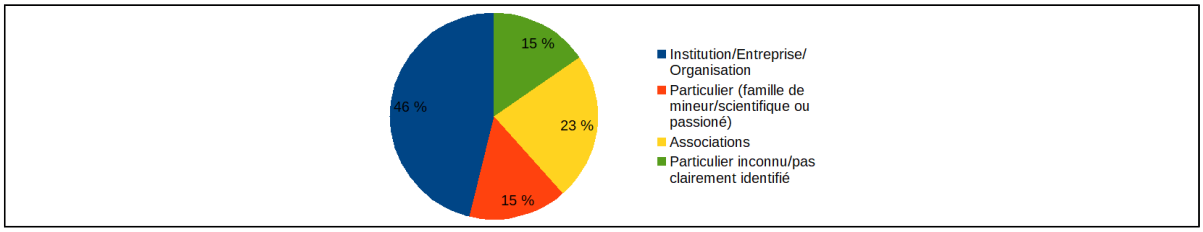

Figure 3 : Répartition quantitative et typologie des acteurs des sites web/Internet de la Figure 2

ENTRÉE	SOURCES2	Définition	EQUIVALENT FRANÇAIS	SYNONYM	SYNONYME	RENVOIS /	ANTONYM	VARIANTE
ABANDER	memo-mines_RL_11	Boiser provisoirement pour a	(vide)	(vide)	ABINDER	(vide)	(vide)	(vide)
ABATTAGE	memo-mines_RL_12	Action de détacher le charbo	(vide)	(vide)	ABATAGE	(vide)	(vide)	(vide)
ABATTEUX	memo-mines_RL_42	Ou piqueur ou haveur. Ouvri	(vide)	(vide)	PIQUEURHAVEUR	(vide)	(vide)	(vide)
ABLO	memo-mines_RL_11	1 – pièce de bois servant d'a	(vide)	(vide)	HABLOT | ABLOC	(vide)	(vide)	(vide)
AÉRAGE	memo-mines_RL_10	Action de faire circuler de l'a	(vide)	(vide)	VENTILATION	AÉRATION	(vide)	(vide)
AIGUILE	memo-mines_RL_11	1 - long forêt servant à faire	AIGUILLE	(vide)	AIWILLE	AIGUILE-	(vide)	(vide)
ARAIOT	memo-mines_RL_11	(vide)	(vide)	(vide)	ARAIOU | ARAYO	ARREYOU-	(vide)	(vide)
ARBINERR'B	memo-mines_RL_11	Faire reculer un wagonnet, re	(vide)	(vide)	ERBINER	(vide)	(vide)	(vide)
ARCANGEAO	memo-mines_RL_11	Emplacement à double voie,	RECHANGEAGE	(vide)	ARQUANCHAGE	(vide)	(vide)	(vide)
ARDI	memo-mines_RL_11	Outil, coin en acier dont le m	HARDILLON	(vide)	HARDI	(vide)	(vide)	(vide)
ARDOUBLAO	memo-mines_RL_11	Prolongement de la journée d	REDOUBLAGE	(vide)	R'DOUBLACHE	(vide)	(vide)	(vide)
ARDOUBLER	memo-mines_RL_11	Action de faire un ardoublach	REDOUBLER	(vide)	R'DOUBLER	(vide)	(vide)	(vide)
AREINE	memo-mines_RL_22	Nom utilisé dans les province	(vide)	(vide)	ARÈNE	(vide)	(vide)	(vide)
ARLET	memo-mines_RL_11	Accident de terrain sur une c	(vide)	(vide)	RELAIS	ACCIDINT D'	(vide)	(vide)
ARREYOU	memo-mines_RL_11	Tige de fer solide, à poignée	(vide)	(vide)	ARAIOTARAIOUA	(vide)	(vide)	(vide)
ARSAQUER	memo-mines_RL_11	Retirer un bois dans une taill	RETIRER	(vide)	R'SAQUER	(vide)	(vide)	(vide)
ASTIQUETTE	memo-mines_RL_11	1 - crochet que le mineur enf	(vide)	(vide)	ESTIQUETTE	(vide)	(vide)	(vide)
BANDE	memo-mines_RL_11	Convoyeur à bande utilisé por	(vide)	(vide)	BANTE	(vide)	(vide)	(vide)
	memo-mines_RL_14	Bande de convoyeur, généra	(vide)	(vide)	TAPIS	(vide)	(vide)	(vide)
BANDE TRAN	memo-mines_RL_10	Système de manutention con	(vide)	(vide)	TRANSPORTEUSE	(vide)	(vide)	(vide)
BARETTE	memo-mines_RL_19	Casque d'un mineur au dépar	(vide)	(vide)	BARRETTE	(vide)	(vide)	(vide)
BARITEL	memo-mines_RL_22	Synonyme de "machine à mo	(vide)	(vide)	MACHINE À MOLE	(vide)	(vide)	(vide)
BARROU	memo-mines_RL_11	Wagonnet vide dans le pas-d	BARROUD	(vide)	BAROU, BAROUT	BALLE-BERL	(vide)	(vide)
BENNE	memo-mines_RL_13	Caisson sur roues roulant su	(vide)	(vide)	CHARIOT BERLIN	(vide)	(vide)	(vide)
	memo-mines_RL_28	Caisson sur roues roulant su	(vide)	(vide)	CHARIOT -- BERL	(vide)	(vide)	(vide)

Figure 4 : Table dynamique permettant l'analyse des *lexical soft data*

Suite à une première analyse terminologique des *lexical soft data,* 37 catégories sur 107 [8] catégories du format *TermBase eXchange* ont été sélectionnées comme potentiellement utiles pour décrire les données. Parmi celles-ci, la catégorie *termType* [9] (*cf.* Figure 5) contient la plupart des valeurs utiles pour une terminologie basique. Si beaucoup des 37 catégories sont actuellement vides, elles ont été sélectionnées en vue d'être enrichie par les *soft data* (*cf.* partie 3.1) et par des entretiens avec des experts.

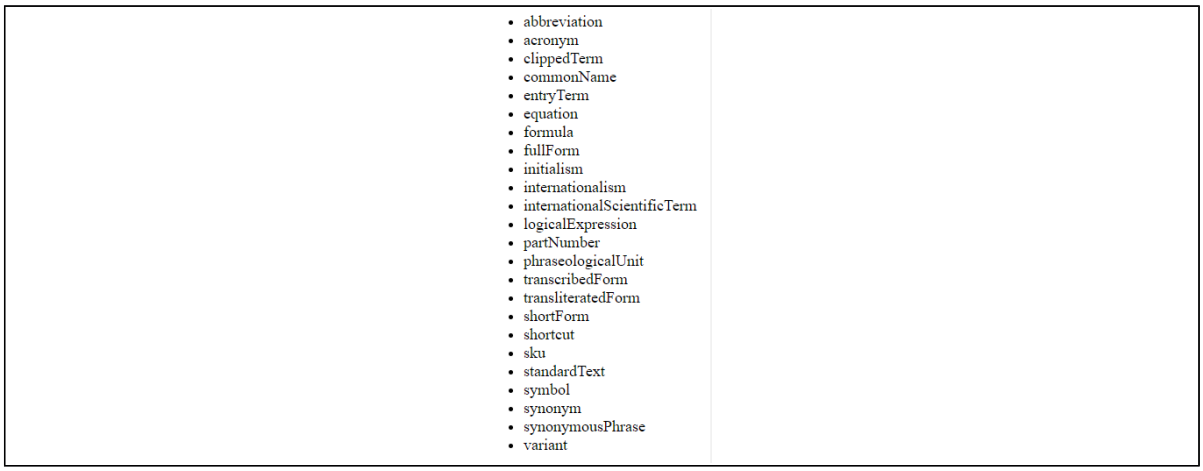

Figure 5 : Valeurs de la catégorie TBX : TermType

ENTRÉE	SOURCES2	Définition	EQUIVALENT FRANÇAIS
RACCOMMODEU	memo-mines_RL_39	Ouvrier chargé de l'	(vide)
RACCOMMODEUR	memo-mines_RL_16	Travaille surtout à l'	(vide)
	memo-mines_RL_18	Vieil ouvrier chargé	(vide)
	memo-mines_RL_19	Ouvrier qui rectifie l'	(vide)
	memo-mines_RL_21	Ouvrier expérimenté	(vide)
	memo-mines_RL_35	Ouvrier expérimenté	(vide)
	memo-mines_RL_56	Vieil ouvrier chargé	(vide)
RACCOMMODEUX	memo-mines_RL_11	Mineur expérimenté	RACCOMODEUR
RACCOMMODEUX D'MINEURS	memo-mines_RL_11	Médecin chargé de	(vide)
RACCOMODEU	memo-mines_RL_7	Vieux mineur	(vide)
RACCOMODEUR	memo-mines_RL_20	Ouvrier chargé de l'	(vide)
	memo-mines_RL_25	Ouvrier chargé de l'	(vide)
	memo-mines_RL_27	Ouvrier chargé de l'	(vide)
RACCOMODEUX	memo-mines_RL_42	Vieil ouvrier dont la	(vide)

Figure 6 : raccom(m)odeu/raccom(m)odeux/raccomodeur

Pour être enregistrées, les *lexical soft data* doivent ensuite être lemmatisées et dédoublonnées. Une des difficultés réside dans la lemmatisation des lexèmes en dialecte (les lexiques contiennent en effet des variantes en chti ou rouchi [10] et ne sont parfois pas uniformisés). Pour la plupart, nous prenons le lemme en français s'il existe (*cf.* Figure 6) tout en gardant les variantes, sinon nous conservons la forme du terme comme il est présenté si aucun autre terme ne s'apparente à un lemme dans son entourage. Seulement 306 termes en dialecte ont un équivalent français.

Dans un deuxième temps, un travail de mise en relation des synonymes doit être réalisé. Si la plupart des *lexical soft data* contiennent des synonymes, la typographie utilisée pour les signaler n'est pas uniforme. Un travail de repérage des expressions marquant celle-ci est effectué. Ainsi, les marques textuelles intra-définitions « ou », « on dit aussi » ou à la fin des définitions « syn. », « synonyme » *etc.* permettent de créer des algorithmes de recherche pour repérer ceux-ci. En revanche, ces algorithmes ne fonctionnent pas infailliblement et une vérification est alors obligatoire :

Exemples pris du blog d'André Paillart [11], un passionné de la mine :

(1) « En photo : un déclimètre ***ou*** *éclimètre et une clé à bidules »*

(2) *Abattage : travail consistant à extraire le charbon de la veine. L'abattage se fait au pic* ***ou*** *au marteau piqueur.*

En (1), la description permet de mettre une relation synonymique « déclimètre » et « éclimètre ». En revanche, en (2), le « pic » n'est pas synonyme de « marteau-piqueur ». La définition nous apprend par contre que l'action d'abattre peut être effectuée par deux outils différents ; un pic ou un marteau-piqueur. Elle permet ainsi de mettre en relation trois termes pouvant être associés.

3.2 …aux *smart data* : étude qualitative pour la mise en correspondance de concepts

L'adjectif *smart* fait, comme nous l'avons vu précédemment, référence à l'aspect structuré des données. L'objectif est de transformer l'ensemble des *lexical soft data* en un SOC raisonné c'est-à-dire un thésaurus au format SKOS explicitant le sens des relations entre les termes retenus. Nous présentons d'une part la méthode d'identification des concepts présents dans la terminologie spécialisée (résultat de l'analyse des *lexical soft data*) et d'autre part, ceux présents dans les outils classificatoires qui organisent les ressources de différentes institutions du domaine.

3.2.1 Identification des concepts des *lexical soft data*

Nous avons évoqué plus haut (partie 3.1) et dans d'autres travaux (Daloz, 2018 ; Daloz et Chaudiron, 2019) les correspondances synonymiques et associatives des termes qui sont différentes en lexicographie et en organisation des connaissances. En ce qui concerne la généricité entre les termes, nous nous inspirons des travaux de Gheorghita (2011 et 2014), qui se base sur l'analyse des définitions du *Trésor de la langue Française informatisé* pour l'indexation et la recherche d'images. Nous utilisons en revanche une méthode plus manuelle et qualitative (Daloz et Chaudiron, 2019).

L'analyse de la généricité entre les termes permet une première conceptualisation du domaine et permet de remarquer que nos *lexical soft data* ne font pour la plupart référence qu'au vocabulaire lié au métier de l'exploitation minière (cf. Figure 7).

outils	équipements de sécurité
transports	dangers
lieux	accessoires d'éclairage
métiers	vêtements
méthodes d'exploitation	matériel
voies de circulation, d'exploitation et de communication	tâches/actions

Figure 7 : termes génériques principaux des soft lexical data

L'analyse des concepts contenus dans les outils classificatoires du domaine permet d'être plus exhaustif.

3.2.2 Identification des concepts des outils classificatoires

Nous appelons « outils classificatoires » ou plus précisément « structures classificatoires » tout système d'organisation des connaissances tels que : terminologie, plan de classement, index thématique, classification qui organise, indexe, structure des documents sur le domaine minier (que ce soit des *soft data* de tous types ou des documents physiques dans les centres de documentation). Nous définissons ceux-ci plus en détail dans notre article (Daloz et Chaudiron 2019) où nous expliquons la méthode d'analyse des thématiques contenues dans ceux-ci qui permet d'aboutir à une première structuration des termes de la terminologie minière. Trois grands domaines ontologiques sont identifiés : le travail à la mine, les mineurs et l'après-mine.

3.2.3 Mise en correspondance des catégories de concepts

La mise en correspondance des deux catégories de concepts (partie 1.3.2.a et partie 1.3.2.b) se base sur une troisième analyse qualitative de quatre ressources : la terminologie minière, le thésaurus Motbis, le thésaurus de l'Unesco et le thésaurus du centre de documentation de Blegny-Mine, sélectionnées pour les raisons énoncées ci-dessous :

- La terminologie minière est le résultat de l'étude des *lexical soft data* enrichie par l'analyse de lexiques (non numérisés) provenant du Centre Historique minier de Lewarde. Elle est importante pour instancier les futurs thésaurus et ontologie et comme nous l'avons vu ci-dessus, pour structurer une partie du domaine.

- Le thésaurus Motbis [12] est un langage contrôlé utilisé pour indexer, échanger et rechercher l'information éducative. Sa structure fait référence à des concepts aussi bien très généraux que très spécifiques et est riche de concepts sur la transmission de savoirs, utiles pour indexer des documents sur le patrimoine minier.

- Le thésaurus de l'Unesco est publié sous *open-access* sous la forme d'une liste de termes contrôlés et structurés pour l'analyse thématique et la recherche de documents et publications dans les domaines de l'éducation, la culture, les sciences naturelles, les sciences sociales et humaines, la communication et l'information. Il complète en cela le thésaurus Motbis et est plus précis sur l'indexation des éléments du patrimoine culturel même s'il contient peu de termes spécifiques au patrimoine minier.

- Le thésaurus de Blegny-Mine indexe quant à lui des ressources documentaires du centre de documentation du site minier de Blegny-Mine en Belgique. Il est construit sur la base de la liste d'autorités RAMEAU et de l'indexation INICHAR (Institut national de l'industrie charbonnière, homologue du CERCHAR [13] en France). Ce dernier est le plus proche du domaine mais nous n'avons pas accès aux différentes relations entre les termes utilisés et ne concerne pas que la mémoire minière.

Nous présentons en Annexe 1 l'ébauche du thésaurus d'après l'analyse des ressources précédentes sous la forme d'une liste thématique (exportée à partir du logiciel *Tematres*).

3.3 Un outil qui facilite la formalisation des données : *Tematres*

3.3.1 Fonctionnalités de *Tematres*

L'outil *Tematres* est un logiciel *open source* qui permet de gérer et d'exploiter des vocabulaires contrôlés, thésaurus ou taxonomies. Il permet notamment de manipuler aisément les données et de les convertir au format SKOS (cf. Figure 8) pour permettre leur ouverture (*open smart data*).

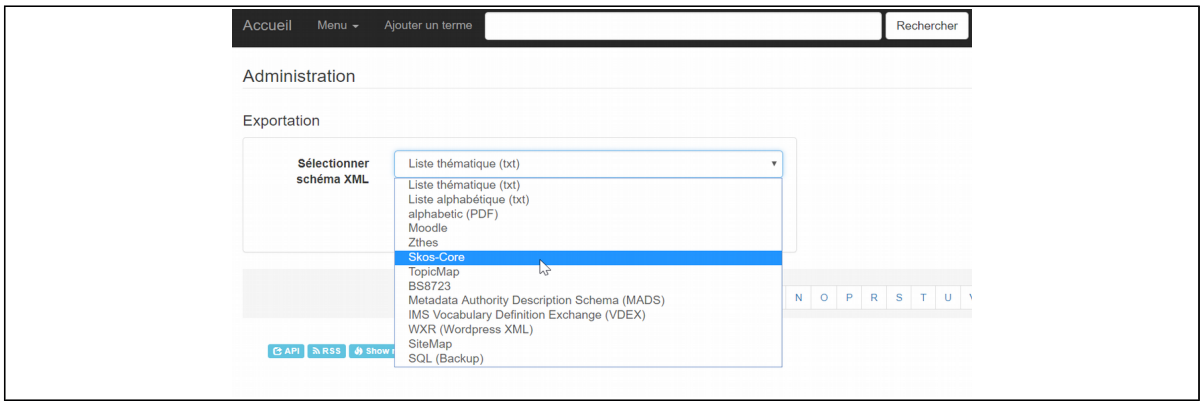

Figure 8 : Différents formats d'exportation du thésaurus sous *TemaTres*

Figure 9 : Meta-term

La Figure 9 représente une partie des méta-termes utilisés qui sont définis dans *Tematres* comme suit : « *A Meta-term is a term that can't be use in indexing process. Is a term to describe others terms. Ej: Guide terms, Facets, Categories, etc.* ». Ils correspondent au premier niveau d'accès au thésaurus et peuvent être de nature ontologique (micro-thésaurus de Motbis ou domaines du thésaurus UNESCO).

Dans ses fonctionnalités, *Tematres* est d'autant plus intéressant qu'il permet l'importation ou la référence à d'autres thésaurus (*cf.* Figure 10), la caractérisation du type de termes spécifiques (partitifs (partie d'un tout) ou instances caractérisés par la relation « *is-a* ») (*cf.* Figures 11 et 12), la possibilité d'ouvrir à plusieurs utilisateurs, la caractérisation des langues, l'ajout de note d'application (historique, bibliographique, définition…) ou pour les termes exclus, la caractérisation des variantes (prononciation, mal orthographiée, abréviation, forme complète…*etc*.).

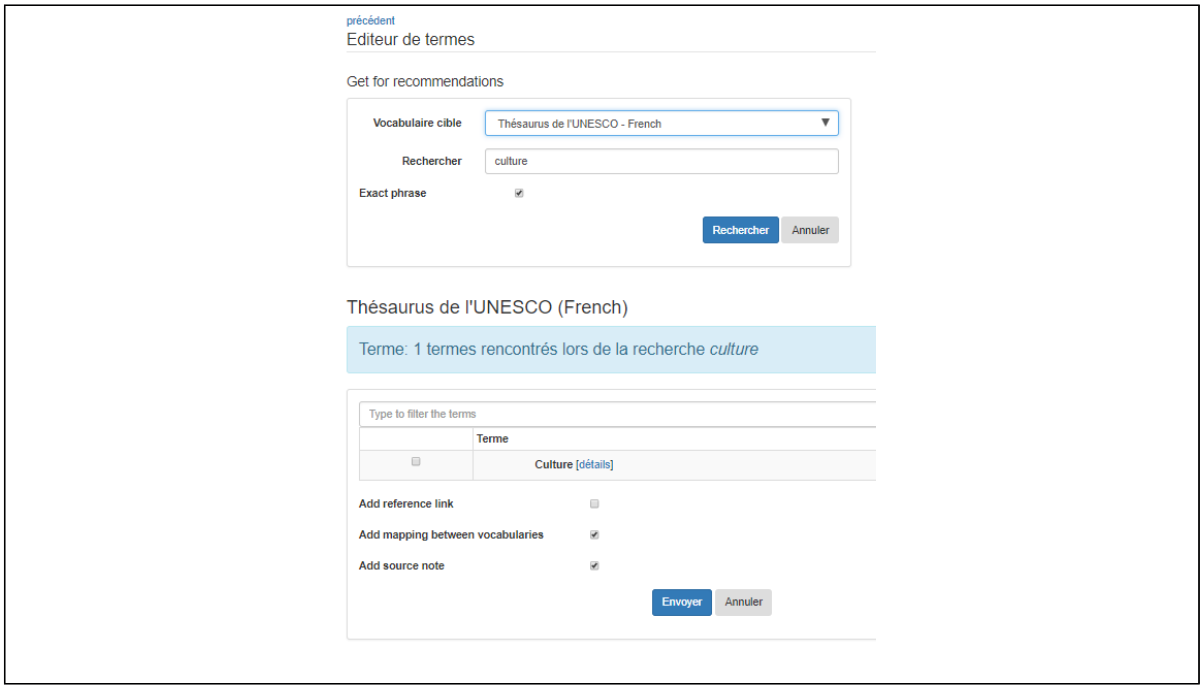

Figure 10 : Importer un concept / insérer un lien vers un descripteur d'un thésaurus externe

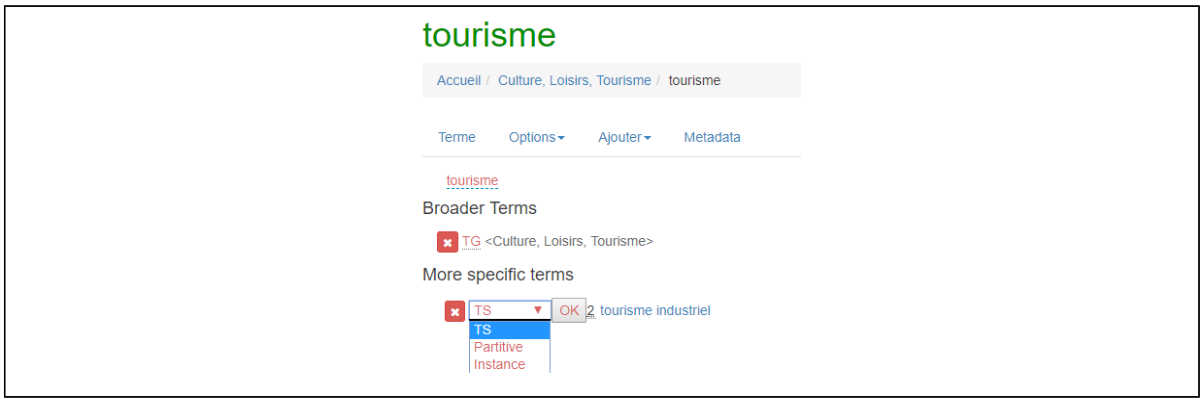

Figure 11 : Terme spécifique partitif ou instance

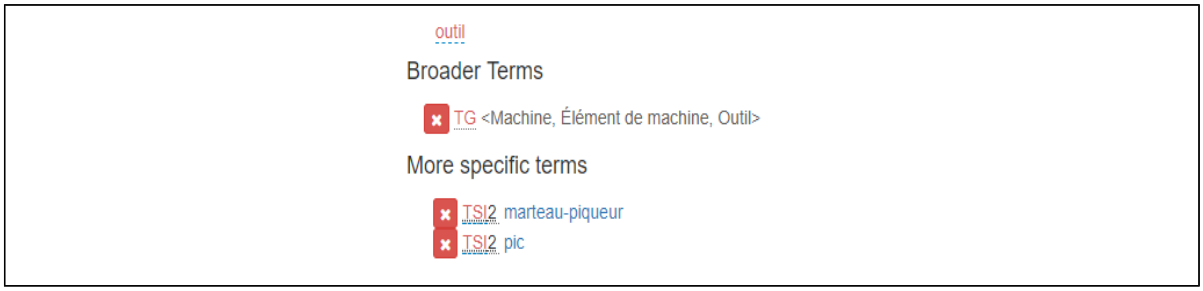

Figure 12 : TSI = Terme spécifique instance

3.3.2 Format SKOS

La formalisation en SKOS permet d'attribuer des identifiants uniques aux différents concepts (URI), identifiables par la chaîne de caractère « tema=*nombre* » dans la Figure 13, qui permettent l'échange des données. Elle caractérise également les relations entre les termes par les balises « skos :prefLabel » pour le candidat descripteur, « skos :broader » pour le terme spécifique, « skos :narrower » pour le terme générique *etc*.

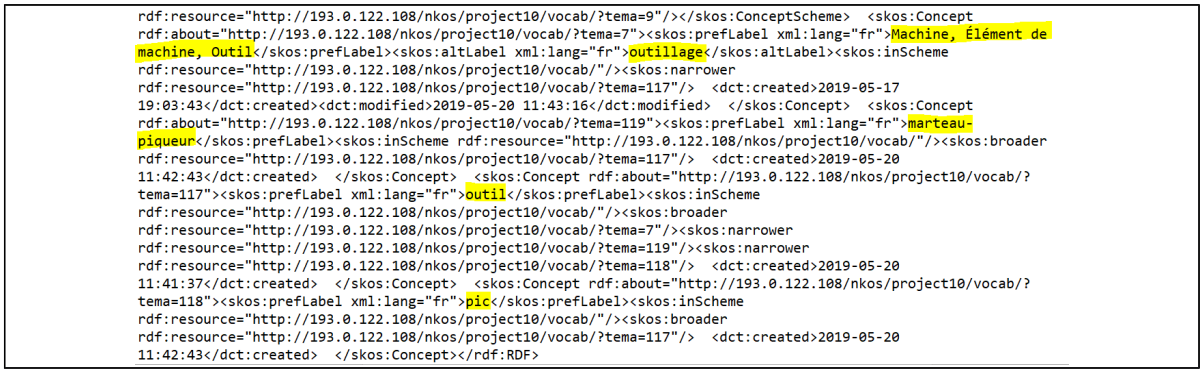

Figure 13 : Capture d'écran d'une partie du fichier en SKOS : Meta-term : machine, élément de machine, outil / TS outil

CONCLUSION ET PERSPECTIVES

Dans cet article, nous avons présenté une double méthodologie fondée d'une part sur la constitution et l'exploitation d'un corpus lexical sur la base de *lexical soft data*, données lexicales du Web 2.0 définissant un domaine et d'autre part, sur l'étude qualitative pour la

mise en correspondance de deux catégories de concepts : d'une part, les concepts présents dans la terminologie spécialisée et d'autre part, ceux présents dans les outils classificatoires qui organisent les ressources de différentes institutions du domaine. Nous avons présenté les résultats obtenus concernant le passage des *lexical soft data* aux *smart data* avec une ébauche du thésaurus en format SKOS.

Les résultats décrits (terminologie et thésaurus) serviront aux professionnels de documentation du domaine (Centre de documentation de Lewarde). Ils pourront également servir de base à certains projets comme le « centre de ressources numériques » de la Mission Bassin Minier à Oignies ou alimentera les bases de données du BRGM [14].

Enfin, comme les travaux de Castéret et Larché (2015, p. 151), notre recherche se place dans le cadre d'une approche sémantique, dont l'intérêt « *réside dans sa capacité à transcrire, à faire éprouver les spécificités du patrimoine culturel immatériel* » grâce notamment à « *l'interopérabilité des bases de données et la nouvelle démarche collaborative permise par la mise en relation de contenus statiques avec des données ouvertes (open data), en ne sélectionnant dans ces dernières que les données intelligentes (smart data) pour les lier au contenu statique* ».

L'objectif final de notre travail de thèse est de créer une base de connaissance compréhensible par les machines. Pour cela, il faut structurer les connaissances et donc créer une ontologie. La démarche mixte présentée dans cet article aidera à sélectionner les concepts d'un modèle ontologique existant pour construire une ontologie de domaine. Adaptée à la définition d'informations muséographiques, le modèle CIDOC CRM nous aidera en effet à structurer les connaissances du patrimoine minier dans le but de valoriser le patrimoine minier.

FINANCEMENT ANR « MEMO-MINES »

Les recherches présentées dans cet article sont financées par le projet ANR-16-CE38-0001 « MEMO-MINES ».

NOTES

[1] Conversion des traces mémorielles en médiations numériques : le cas de la mémoire minière (projet ANR-16-1 CE38-0001-02).

[2] Parmi ces institutions, nous pouvons citer le Centre Historique Minier de Lewarde (CHM), les Archives Nationales du Monde du Travail de Roubaix (ANMT), le Bureau de Recherches Géologiques et Minières à Billy-Montigny (BRGM) ou la Mission Bassin Minier à Oignies.

[3] Par exemple, des membres d'associations, des anciens mineurs ou des familles d'anciens mineurs, des amateurs ou des collectionneurs d'objets

[4] https://www.vocabularyserver.com

[5] « *Furthermore it could be argued that humanities computing is mainly interested in digitalized texts […] and not material that is natively digital. Born digital material would include computer games, blogs, virtual worlds, social spaces […], email collections, websites, surveillance footage, […] and digital art. Most of these « objects » are studied and analyzed within different kinds of new media settings and to me this is an interesting in-between zone. Would humanities computing be interested in engaging more with new media scholars? There is certainly a need for well-crafted tools for studying online life and culture.* »

[6] https://www.w3.org/2004/02/skos/. Consulté le 22/08/2019

[7] https://tectoniq.meshs.fr

[8] http://datcatinfo.net/subset_repos/tbx_master_list

[9] http://www.datcatinfo.net/datcat/DC-2677

[10] Le rouchi est une variété locale du picard en usage dans la région de Valenciennes.

[11] https://andredemarles.skyrock.com/3113954741-Lexique-des-termes-miniers-1ere-partie-de-A-a-D.html

[12] http://www.cndp.fr/motbis/

[13] Centre d'Etudes et de Recherches des Charbonnages de France

[14] Bureau de Recherches Géologiques et Minières

RÉFÉRENCES BIBLIOGRAPHIQUES

CASTÉRET Jean-Jacques, LARCHE Mélanie (2015). Le projet « PCILAB » pour la valorisation numérique de l'inventaire français du PCI. In : SEVERO Marta, ROMELE Alberto. (2016). *Traces numériques et territoires*. Paris : Mines-ParisTech, 2015, 270p. (Territoires numériques).

DALOZ Amélie (2018). Vers la représentation terminologique du patrimoine minier. *Terminologie & Ontologie : Théories et application : actes de Colloque TOTh*, Actes à paraître.

MEHTA Dinesh P, SAHNI Sartaj (eds) (2005). *Handbook of Data Structures and Applications.* Boca Raton : Chapman & Hall, 2004, 1392p.

DALOZ Amélie, CHAUDIRON Stéphane (2019). Méthodologie de structuration d'un thésaurus du domaine minier ». In : JACQUEMIN Bernard et GHENIMA Malek. *La numérisation info-documentaire : actes de Colloque International sur le Document Numérique*. Paris : Europia productions, 2019, p.11-21.

GHEORGHITA Inga (2011). Méthodologie de construction automatique du thesaurus pour l'indexation et la recherche des images*. Actes de la 13e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'2011*) [en ligne]*, 2011.* Disponible sur : https://hal.archives-ouvertes.fr/hal-00695722/document

GHEORGHITA Inga (2014). *Construction automatique de hiérarchies sémantiques à partir du Trésor de la Langue Française informatisé (TLFi) : application à l'indexation et la recherche d'images.* Sciences du langage. Nancy : Université de Lorraine, 2014, 240p.

OCDE (2007). *Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics.* OCDE, 2007, 29p. Disponible sur : http://www.oecd.org/fr/sti/inno/38500823.pdf

SCHÖCH Christof (2013). Big? Smart? Clean? Messy? Data in the Humanities. *Journal of Digital Humanities* [en ligne], 2013, vol. 2 n°3, p. 2-13. Disponible sur : https://hal.archives-ouvertes.fr/hal-00920254

TERRAS Melissa, NYHAN Julianne, VANHOUTTE Edward (2013). *Defining digital humanities*: *a reader.* Farnham, Royaume-Uni : Ashgate, 2013, XV-314p.

UNESCO (2003). *Convention for the Safeguarding of the Intangible Cultural Heritage*. Paris : UNESCO, 2003, 19p. Disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000132540

Annexe

ÉBAUCHE DU THÉSAURUS (LISTE ALPHABÉTIQUE À PARTIR DE L'EXPORTATION *TEMATRES*)

Comportement social, Vie social

. vie quotidienne

Culture, Loisirs, Tourisme

. activité de loisir

. . activité manuelle

. . jardinage

. . activité sportive

. . football

. . activité musicale

. . activité hialeutique

. . activité ornithologique

. . . colombophilie

. . . sérinophilie

. art vivant

. . théâtre

. . danse

. . chanson

. fête et anniversaire

. littérature

. . poésie

. tourisme

. vacances

Energie, Ressources naturelles

. ressource énergétique

. . carburant

. . . gaz naturel

. . . . gaz de schiste

Enseignement

. école

. formation

Entreprises

. organisation

. organisation patronale

. organisation syndicale

. personne morale

Géographie, Géologie

. cartographie

. gisement

Habitat, Urbanisme

. agglomération urbaine

. habitat

. logement (habitation)

. urbanisme

Immigration

. langue

. population

. racisme

Lieux et espaces aménagés

. génie civil

. partie industrielle

Nature et environnement

. sciences naturelles

. . faune

. . flore

Recherche

. ethnographie

. géographie

. langage

. sociologie

Religion et croyances

. chant religieux

. édifice religieux

. lieu de sépulture

. office religieux

. procession

. rite

Santé

. équipement sanitaire

. hygiène

. . santé publique

. . . hygiène de l'habitat

. . . hygiène professionnelle

. maladie professionnelle

. . silicose

. . surdité

Sécurité

. accident

. . catastrophe

. . coup d'eau

. . coup de grisou

. . coup de poussière

. . éboulement

. équipement de sécurité

. . casque

. . vêtement de travail

Techniques d'exploitation, Technologie

. hydraulique

. pneumatique