

# Analyse de corpus de presse pour une valorisation du patrimoine minier en France : projet ANR MémoMines

Amara Cozzarolo

▶ **To cite this version:**

Amara Cozzarolo. Analyse de corpus de presse pour une valorisation du patrimoine minier en France : projet ANR MémoMines. Sciences de l'information et de la communication. 2021. dumas-03343093

**HAL Id: dumas-03343093**

**https://dumas.ccsd.cnrs.fr/dumas-03343093**

Submitted on 13 Sep 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**Amara COZZAROLO**

**Master Information Documentation**
**Première année**

**MEMOIRE DE STAGE**

**Mission effectuée du 03/05/2021 au 05/07/2021**

**au**
**Laboratoire GERiiCO**
**Villeneuve d'Ascq**

# Analyse de corpus de presse pour une valorisation du patrimoine minier en France : Projet ANR MémoMines

**Sous la direction de :**

**M. B. JACQUEMIN (tuteur universitaire)**

**M. E. KERGOSIEN (tuteur professionnel)**

**Soutenu le xx Juin 2021 à l'UFR DECCID-SID**
**Université de Lille, Sciences Sociales et Humaines**
**BP 60 149, 59 653 Villeneuve d'Ascq Cedex**

**Année Universitaire 2020/2021**

*Je remercie tout d'abord, **Monsieur CHAUDIRON**, **Monsieur Kergosien** et **Monsieur JACQUEMIN** de m'avoir accueillie au sein de ce projet, de m'avoir encadrée et aidée durant ces mois de stage et de m'avoir transmis leurs savoirs sur ce domaine passionnant.*

*Je remercie évidemment toute **l'équipe pédagogique du Master 1 Information-Documentation**, pour leur soutien et leur compréhension face à cette période de crise sanitaire.*

*Je remercie également mes camarades, **Arezki ACHOURI**, **Alexandre RIOTTE** et **Cerine BOUBEHA** qui m'ont épaulé, rassuré et soutenu durant toute cette année, ainsi que **Ghita JOUNDI** qui s'est toujours rendue disponible pour m'écouter.*

*Enfin, je souhaiterais aussi remercier **Romain DEMARTHE** pour son soutien infaillible, et ce, depuis le début de ce cursus.*

# Résumé et mots-clés

La valorisation du patrimoine représente une certaine forme d'héritage d'une société qui n'existe plus dans les faits mais dont on se souviendra dans les années à venir. Ce mémoire présente le traitement et l'analyse d'un corpus de presse dans le cadre du projet ANR MémoMines qui a été effectué lors d'une mission de stage au laboratoire GERiiCO de l'Université de Lille. Le but de cette mission étant de proposer une grille d'analyse du contenu permettant de repérer les acteurs et les lieux ayant un lien avec le domaine de la mine. La première étape de cette mission a été de nettoyer le corpus d'articles de presse à l'aide du langage XSLT qui est un langage de structuration de l'information. Afin de mener à bien cette mission, une méthodologie a été mise en place, notamment avec l'utilisation du logiciel Tropes, un logiciel d'analyse sémantique. Cette méthodologie a permis de développer plusieurs scénario visant à confirmer des listes préétablies d'acteurs, de lieux ainsi que le thesaurus déjà défini au préalable.

Mots-clés : Patrimoine – Analyse de contenu – Tropes - Modélisation – Structuration de l'information – XSLT

# Abstract and key-words

The heritage enhancement represent a certain form of heritage of a society which no longer exists in practice but which will be remembered for years to come. This thesis present the treatment and analysis of a press corpus in the ANR MémoMines project context which was carried out during an internship mission at the GERiiCO laboratory at the University of Lille. The goal of this mission is to propose a content analysis grid to identify the actors and places having a link with the field of the mine. The first step of this mission was to clean up the press articles corpus using the XSLT langage which is an information structuring langage. In order to carry out this mission, a methodology was put in place, in particular with the use of the Tropes software, a semantic analysis software. This methodology made it possible to develop several scenarios aimes at conforming pre-established lists of actors, places as well as the thesaurus already defined beforehand.

Key-words : Patrimony – Content Analysis – Tropes – Modelization – Information Structuring – XLST

# SOMMAIRE

1 Introduction.....5
1.1 Contexte.....5
1.1.1 Présentation du projet.....5
1.1.2 La position du laboratoire GERiiCO.....5
1.1.3 La mission de stage.....6
1.2 Cadrage du sujet.....6
2 Analyse de contenu avec Tropes.....8
2.1 Choix des outils.....9
2.1.1 Présentation des différents outils.....9
2.1.2 Choix de Tropes.....10
2.1.2.1 Fonctionnement de Tropes.....10
2.1.2.2 Pourquoi Tropes ?.....13
2.2 Méthodologie utilisée.....14
3 Structuration de l'information.....22
3.1 Définitions.....22
3.2 Balisage et langages de structuration de l'information.....23
4 Conclusion et perspectives.....27
5 Références bibliographique.....28
5.1 Articles.....28
5.2 Sites web.....28
6 Annexe.....30

# 1 Introduction

## 1.1 Contexte

### 1.1.1 Présentation du projet

Le projet ANR MEMO-MINES a été lancé, à date effective, le 16 janvier 2017 et devra prendre fin fin quatre ans après. Ce projet répond à une problématique de sauvegarde des mémoires individuelles et de conservation de faits ou d'événements qui constituent le patrimoine minier des Hauts-de-France. Le défi est de formaliser l'information, de créer une base de connaissance, car il y a une diversité et une incomplétude des données et des ressources. Le but est de mettre à disposition ses informations sous corpus d'archives numérique et qui sera distribué dans les lieux de mémoire. Cela constitue donc un enjeu social qui est lié à la collecte, l'analyse, la publication et la mise en valeur de l'histoire et de la mémoire collectives des acteurs de la mine que se soit à l'oral ou à l'écrit.

Les objectifs de ce projet sont donc d'apporter les outils et les méthodes scientifiques pour la sauvegarde et la mise en valeur du patrimoine minier des Hauts-de-France et cela doit passer tout d'abord par la constitution et la production de la mémoire minière que se soit des traces mémorielles à numériser, des entretiens avec des anciens professionnels de la mine et la restitution du cadre de vie à l'époque minière. Cela passe ensuite par l'explication de l'univers de connaissance de la mine, la réalisation des archives audiovisuelles de la mémoire minière, la conception et la mise en place de dispositifs techniques innovateurs pour exploiter et remédiatiser et pour finir par la spécification et la mise en place d'une « plateforme de travail ».

Ce projet constitue donc deux enjeux principaux. Le premier est donc un enjeu scientifique et technique puisqu'il faut décrire l'univers du patrimoine minier et le numériser. Le but étant de proposer une meilleure production, traitement, archivage, diffusion et réutilisation de données numériques car ce n'est pas assez approfondi dans la littérature spécialisée. Le second enjeu est un enjeu expérimental et de terrain, le but est de produire un corpus indexé sur le sujet du patrimoine minier et de le rendre accessible par tous, en ligne, et pour être utilisé comme ressources numériques. Ici, on recherche la préservation de l'héritage de la mine et sa valorisation.

Le projet ANR MEMO-MINES s'attaquera tout d'abord à la valorisation d'une mémoire régionale, pour ensuite passer à la mémoire nationale et finir par les autres régions minières comme en Allemagne par exemple. Les ressources intellectuelles qui en ressortiront serviront pour des projet de médiation culturelle, de valorisation territoriale et de recherche.

### 1.1.2 La position du laboratoire GERiiCO

Le laboratoire GERiiCO est un laboratoire de recherche internationale en science de l'information et de la communication. Il est situé dans la région des Hauts-de-France à l'université de Lille. Il est constitué de 38 enseignants-chercheurs de l'université de Lille qui travaillent principalement sur l'analyse des pratiques, des processus et des dispositifs info-communicationnels.

Dans ce projet, GERiiCO doit s'occuper du recueil ainsi que de l'organisation et la gestion des connaissances minières mais aussi de mettre en œuvre et expérimenter l'exploitation experte de l'archive. Les enjeux étant de définir un modèle de représentation et d'organisation des ressources numériques pour qu'elles puissent être utilisées de façon idéale, de réaliser l'ontologie du domaine à partir du recueil des connaissances expertes et des pratiques amateurs afin d'analyser les corpus collectés en modèle de description ainsi que des thésaurus spécifiques. Aussi, GERiiCO va devoir imager, concevoir et réaliser des scénarios spécifiques de l'archive afin de la mettre à disposition d'un public expert et à un environnement de publication. Ce dernier enjeu répond au besoin de remédiatisation des données.

### 1.1.3 La mission de stage

Ce stage s'effectuera au laboratoire GERiiCO dans le cadre du projet MEMO-MINES. Dans ce projet, un des objectifs est d'analyser un corpus d'articles de presse provenant du journal *La Voix du Nord* et portant sur le domaine de la mine.

La mission principale du stage est de compléter et d'analyser le corpus de presse grâce au serveur Europresse et à l'outil Tropes afin de mettre en place une grille d'analyse de contenu et de repérer les acteurs qui ont un lien avec la mine. En d'autres termes, le but est de procéder au marquage des documents pour enrichir les ressources et avoir une vision plus précise de ce qu'il y a comme données dans la presse entre 2004 et 2021.

Très concrètement, les tâches que l'on m'a confié ont été tout d'abord de compléter le corpus d'articles de presse avec l'aide du serveur Europresse. Il a fallut ensuite nettoyer le corpus qui comportait trop de bruit (des pages de garde, des sommaires, des numéros de page, des certificats d'utilisation, etc.) pour l'analyser. Il ne fallait garder que les titres, les contenus, les dates de publication, le lieu de publication et le numéro des articles. Dans cette tâche, le langage XSLT a été très utile pour transformer le corpus, qui était sous format HTML, vers des formats XML, HTML et TXT.

Après le nettoyage et la préparation du corpus pour l'analyse, le thésaurus ThésoMines, effectué par Mr Stéphane Chaudiron au préalable, devait être testé à l'aide de l'outil Tropes. La tâche suivante a été d'analyser le corpus d'articles de presse avec le thésaurus, mais aussi aux scénarios de lieux et d'acteurs dont les listes étaient déjà établies avant le début de ce stage. Cette tâche m'a mené jusqu'à la manipulation de Gephi, un outil de cartographie.

## 1.2 Cadrage du sujet

Afin de cadrer correctement notre sujet, nous allons commencer par définir les notions de « patrimoine minier », de « valorisation de patrimoine » et de « base de connaissance ». Nous présenterons ensuite notre problématique et nous finirons par annoncer le plan.

Le patrimoine minier dans ce projet de recherche représente tout le patrimoine culturel matériel, immatériel et langagier du bassin minier des départements du Nord et du Pas-de-Calais. Il désigne tout le patrimoine généré par les activités minières du Hauts-de-France. Dans ce patrimoine matériel on retrouve tous les sites miniers, tout ce qui a été bâti et fabriqué pour servir les acteurs. Le patrimoine immatériel sont toutes les pratiques sociales et culturelles conçues par cette société minière. En d'autres termes, il consiste à rassembler plus de deux siècles d'exploitation du charbon sur un territoire qui a accueilli des dizaines de nationalités ce qui a donc fait cohabiter énormément de langues et de variations de dialecte pendant l'exploitation mais aussi après la fermeture de la dernière houille. Ce sont ces différences de langues et de dialectes qui ont façonné la culture minière. Le patrimoine minier fait parti des biens inscrits au patrimoine mondial de l'UNESCO depuis 2012.

La valorisation d'un patrimoine est un processus qui a pour but d'améliorer la valeur d'un patrimoine. Le patrimoine « *repose avant tout sur la notion de transmission et non simplement sur celle de la possession. La détention du patrimoine par les générations est illusoires, car en réalité c'est un prêt : il y nécessité de rendre dans les meilleures conditions possibles le patrimoine aux générations futures* » (Leroux, 2014). La valorisation d'un patrimoine est un processus très lié à la notion de restitution et de préservation ce qui va créer cette forme de mémoire étant donné qu'elle restitue les références collectives partagées par des individus dans une société donnée. La valorisation d'un patrimoine, de par sa restitution et la préservation de ce dernier, représente une certaine forme d'héritage d'une société qui n'existe plus dans les faits mais dont on se souviendra dans les années à venir. En plus de cette forme d'héritage, elle constitue un véritable atout pour le tourisme d'un pays ou du moins, dans ce cas de figure, d'une région.

Une base de connaissance est, dans le projet ANR MémoMines, la finalité de la valorisation du patrimoine minier des Hauts-de-France. En effet, cette base de connaissance, ici numérique, regroupera toutes les connaissances spécifiques du domaine de la mine dans le but d'être exploitée par le plus grand nombre.

Ces connaissances pourraient correspondre aux langages qu'utilisaient les individus de cette société minière au moment de l'exploitation du charbon mais aussi ces dialectes qui nous sont restés dans notre langage d'aujourd'hui. Cela pourraient aussi montrer tous les savoirs-faire de cette société, ses idées et ses concepts.

Maintenant que les notions principales de ce sujet ont été définies, nous allons observer, dans le cadre de la recherche, comment l'analyse de corpus de presse s'intègre dans un projet de valorisation du patrimoine minier ? Quels sont les outils qui permettent d'automatiser cette mission et pourquoi avoir choisi Tropes plus particulièrement ? Quels sont les critères de sélection pour la collecte d'articles de presse ? Et pour finir, quel est le rôle de la structuration de l'information dans ce processus ?

Afin de répondre à cette problématique, nous allons traiter le sujet de l'analyse de corpus avec l'outil Tropes, que nous définiront afin de cadrer correctement cette partie. Nous allons établir la méthodologie qui sera utilisée dans le cadre d'un stage dans un projet de recherche tel que le projet ANR MémoMines et nous nous attarderons sur la notion d'objectivité qui est une notion importante dans l'analyse de contenu. Nous traiterons ensuite la structuration de l'information et sa place dans la conception d'une base de connaissance à une visée de valorisation d'un patrimoine. Nous commencerons, dans cette partie, par définir très rapidement certaines notions telles que celles d' « information », de « document », de « données » et de « signal ». Puis nous continuerons sur la méthode de balisage et terminerons par voir quels sont les langages qui pourraient nous être utiles dans ce processus.

# 2 Analyse de contenu avec Tropes

L'analyse de contenu est une méthode d'étude scientifique dans les Sciences Humaines et Sociales. C'est un champ de pratique large avec des méthodes, des méthodologies et des techniques scientifiques diverses. Elle est utilisé en recherche mais plus particulièrement dans les domaines des Sciences de l'Information et de la Communication (SIC) et des Sciences de l'Information et du Document (SID). L'analyse de contenu « *se définit comme une technique permettant l'examen méthodique, systématique, objectif et, à l'occasion, quantitatif du contenu de certains textes en vue d'en classer et d'en interpréter les éléments constitutifs, qui ne sont pas totalement accessibles à la lecture naïve* » (Robert, Bouillaguet, 2007). Le terme de « texte » peut désigner aussi bien des productions orales ou écrites, textuelles ou sonores. Dans notre cas, nous nous focaliserons sur les textes d'articles de presse écrite.

Dans l'analyste de corpus de presse, la posture de l'analyste doit être différente de la posture du lecteur de journal. Pour le projet ANR MémoMines, l'analyste doit étudier les numéros du journal papier La Voix du Nord téléchargé sous format HTML, donc sous format numérique. Ce qu'il faut comprendre dans l'analyse de corpus de presse est que les textes qui ont été donnés à étudier pour l'analyste ne sont uniquement les textes que le journal a donné à lire. De plus « *ce que lit l'analyste n'est pas ce qu'à lu le lecteur, ce que lit l'analyste n'est pas ce qui a été dit. […] ce ne sont pas des relations d'identité et de textualité qui unissent les paroles des commentateurs s'exprimant dans l'espace public et le journal qui les fait parler sous son nom, mais des rapports de sélection et de transformation* » (Krieg, 2000). Il est important de préciser aussi que le rôle de l'analyste n'est pas d'interpréter avec son imagination. Aussi, « *ce qui est écrit dans le journal n'est pas ce qui a été dit* » (Krieg, 2000). En d'autres termes, les textes ne reflètent absolument pas des paroles et doit être pris de façon autonome. L'analyse de corpus de presse ou de discours de presse est une spécificité dans l'analyse de contenu. Elle possède ses propres méthodes et techniques scientifiques.

En plus de la spécificité qu'est l'analyse de corpus de presse, il y a une dimension automatique à ajouter. En effet, dans la mission de stage, le but est d'analyse de façon automatisée un corpus de presse. En ce sens, certains outils numériques existent, qu'ils soient gratuits ou payants, et sont dédiés à ce champ disciplinaire. Ces outils permettent de faire ressortir très rapidement les mots-clés d'un discours. Le choix des outils numériques d'analyse sémantique diffèrent en fonction des besoins de l'analyste, la nature des discours à analyser ainsi que leur format. Cependant, ces outils comportent tous leurs propres limites dans leur utilisation. C'est pourquoi, une vue d'ensemble sur les outils existants et leur fonctionnalité va nous être utile pour déterminer le ou les outil(s) qui nous sera(ont) indispensable(s) et qui correspondra(ont) au mieux à nos attentes et nos besoins.

Nous allons voir quels sont les différents outils qui peuvent être utilisés dans le cadre d'une analyse de corpus de presse et pourquoi nous avons choisi Tropes. Nous continuerons sur la méthodologie à adopter, les critères d'analyse et la confection d'une grille d'analyse. Puis nous terminerons par aborder la notion de l'objectivité et sa place dans l'analyse de contenu.

## 2.1 Choix des outils

Dans le cadre du projet ANR MémoMines, qui vise à établir une liste des acteurs et des lieux qui ont joué un rôle pendant ou après la période de l'exploitation du charbon, une question s'est posé sur le meilleur outil à utiliser pour une analyse de données textuelles informatisée. Ils sont nombreux à exister mais il a fallut trouver celui qui répondait le mieux à nos besoins. Pour ce faire, nous allons commencer par présenter certains outils qui pourraient nous intéresser dans le cadre d'une analyse de corpus d'articles de presse automatisée, puis nous allons nous focaliser sur l'outil Tropes, qui est l'outil qui a été choisi pour cette mission de stage.

### 2.1.1 Présentation des différents outils

Avant de commencer quelconque présentation, il est important d'expliquer les données que nous devons analyser. Les données à analyser sont sous forme de corpus textuel. Ce corpus représente l'ensemble des articles du journal La Voix du Nord publiés sur le domaine de la mine entre 2004 et 2021. Il est constitué de plus de 1 300 articles de presse. Ce qu'il faut comprendre c'est que nous avions besoin d'un outil capable de traiter les milliers de pages d'un corpus constitué au préalable et importé par l'analyste dans cet outil. Maintenant que le sujet d'analyse et expliqué, nous pouvons commencer la présentation des cinq outils qui pourraient être utiles dans ces circonstances.

Le premier outil qui sera présenter est Iramuteq. Iramuteq est un logiciel, créer en 2010, qui « *permet de faire des analyses statistiques sur des corpus texte et sur des tableaux. Il repose sur le logiciel R et le langage Python* » (Ratinaud, 2021)[1]. C'est un logiciel disponible sous Windows, Linux et MacOs. Il permet de générer en sortie des corpus sous format CSV et ne prend qu'en entrée uniquement des corpus CSV. L'analyse de données textuelles qu'il propose repose sur la méthode de classification hiérarchique descendante de Reinert (1983, 1991), c'est-à-dire qu'elle est composée de trois modalités : la première étant la classification simple sur texte qui traite les textes dans leur intégralité et qui permet de regrouper les plus proches ; la seconde est une classification qui porte sur les segments de textes ; et la dernière est une classification sur deux tableaux, ici il n'est plus question de segments mais de regroupements de segments. L'un de ses points fort est qu'il propose une richesse informationnelle et une diversité des visualisation puisqu'il associe les différentes classes thématiques à une couleur. De plus, « *IRaMuTeQ s'articule avec TXM : tout corpus qui a été importé dans l'un des logiciels est importable dans l'autre* » (Pincemin, 2018).

Le second outil est TXM qui est un logiciel de textométrie gratuit et open-source, crée en 2009. Ce logiciel est disponible sous Windows, MacOS X et Linux. Il prend en entrée plus d'une dizaine de formats de fichiers et il peut en générer autant en sortie. Il a pour objectif de repenser les calculs textométriques pour des corpus annotés et structurés. Il est utilisé pour l'analyse de données textuelles, du discours, et du contenu ce qui lui permet d'avoir une exploitation plutôt complète des informations contenues dans un corpus. Il propose aussi une personnalisation par des scripts, des sorties graphiques interactifs et les résultats donnés peuvent être exploités par d'autres outils. C'est un logiciel très apprécié dans le domaine des sciences humaines et sociales et plus particulièrement dans les humanités numériques.

Les troisièmes logiciels sont les logiciels Hyperbase et Hyperbase Web Edition qui servent à l'exploration documentaire et statistique des textes. Il a été crée en 1989 mais il a, depuis, évolué et une version 10 en est sortie cette année, en 2021. Il est disponible uniquement sur Windows. Il y a deux fonctions qui sont combinées dans ces logiciels, une fonction documentaire et une fonction statistique. Parmi les fonctions documentaires, on retrouve une lecture naturelle du corpus, une possibilité de navigation par mots-clés, il recherche et tri les contextes et propose un index, des dictionnaires des formes, des lemmes, des codes et des fréquences. Quant à la fonction statistique, elle propose des graphes des unités, des représentations factorielles ou arborées, une extraction des segments du corpus et des représentations des cooccurences

1 Selon le site : http://www.iramuteq.org

et réseaux thématiques[2]. La différence entre Hyperbase et Hyperbase Web Edition est que ce dernier est une refonte du premier dans le but de rendre l'interface plus intuitive et disponible en ligne.

Le quatrième logiciel est Analec. Analec est un logiciel d'analyse et d'annotation de corpus écrits. Il est téléchargeable en ligne gratuitement et disponible sous Windows, Mac et Unix. Il permet la gestion, le regroupement, la visualisation et l'impression de corpus annoté et proposent des calculs de fréquences, de recherche de corrélations et une générations de tableaux qu'il est tout à fait possible de définir soi-même. Il génère en sortie des corpus sous les formats TXT, CSV et GLOZZ.

Le cinquième et dernier logiciel de cette liste se nomme Glozz. C'est un logiciel d'annotation manuelle et d'exploration de corpus textuels. Il est disponible sous OS X, Linux et Windows. Il prend en entrée uniquement des corpus en texte brut et génère en sortie du texte brut ou du CSV. Il permet une annotation manuelle de texte qui peut être déclarée en XML et que l'on peut explorer grâce à des requêtes en langage GlozzQL ou SQL.

Cette liste de logiciels est une liste non exhaustive des logiciels qui pourraient répondre le plus à nos besoins. Ces logiciels sont très intéressants lorsque l'on cherche à analyser du texte et le marquer et pour certains sont très complets en proposant des tableaux et des schémas. Cependant, il a une dimension à prendre en compte dans le choix d'un tel outil, la dimension de proximité.

### 2.1.2 Choix de Tropes

Le logiciel qui a été choisi pour mener à bien la mission d'analyse de corpus d'articles de presse est le logiciel Tropes. Nous allons, dans cette partie, expliquer le fonctionnement de Tropes et pourquoi nous avons opté pour celui-ci.

Tropes est un logiciel d'analyse sémantique. Il a été crée par Pierre Molette et Agnès Landré en 1994. Il est capable d'éditer des ontologies, de proposer une arborescence des références, une analyse chronologique des textes, catégoriser des termes, analyser les acteurs, permettre une extraction terminologique et un diagnostic du style du texte. Il est disponible uniquement sous Windows et c'est un logiciel gratuit. Il est utilisé en science de l'information afin de générer et de vérifier la pertinence des ontologies et des thesaurus. Le but de ce logiciel est de montrer à son utilisateur le squelette du texte analysé.

#### 2.1.2.1 Fonctionnement de Tropes

Comment fonctionne le logiciel Tropes ? Avant de procéder à quelconques manipulations du logiciel, il est important de constituer son corpus, si c'est un corpus qu'il faut analyser, ou son texte. Tropes accepte les entrées sous format texte ANSI, HTML, Microsoft Word, etc. Dans son analyse, Tropes ne prend en compte uniquement les mots et la ponctuation. Les caractères spéciaux ne sont donc pas un problème et n'altère en aucun cas l'analyse sémantique. Cependant, les termes constitués de plusieurs mots doivent être lié avec le caractère « _ » pour compter comme un seul et même mot.

Pour analyser notre corpus et notre texte, il va falloir lancer le logiciel et ouvrir le ou les fichier(s) dans le menu [Fichier] et sélectionner les textes à analyser. Il est tout à fait possible, dans un corpus, de faire en sorte que le logiciel nous montre les résultats d'un seul fichier à la fois. Pour ce faire, il faudra aller sur l'onglet [Fichier] et sélectionner dans la liste des fichiers constituant le corpus, le fichier que l'on souhaite voir uniquement. Il est aussi possible d'en ajouter, d'en effacer ou de les trier.

Les résultats de l'analyse apparaissent sur le côté gauche de la fenêtre principale du logiciel. Afin de mieux comprendre les termes d'analyse, nous allons proposer quelques définitions de ceux-ci. Le premier terme d'analyse est le style. Il précise le style général du texte et sa mise en scène, le nombre de fichiers examinés, le nombre de propositions remarquables qui résument les parties les plus caractéristiques du texte, et le nombre d'épisodes qu'il a pu détecter. Les épisodes sont les parties représentatives du textes et sont constitués selon la chronologie du discours. Dans ces épisodes, on retrouve la notion de rafales, qui sont des mots que l'on retrouve plusieurs fois dans le discours dans un épisode donné.

---

2 Selon la page Wikipédia « Hyperbase », URL : https://fr.wikipedia.org/wiki/Hyperbase

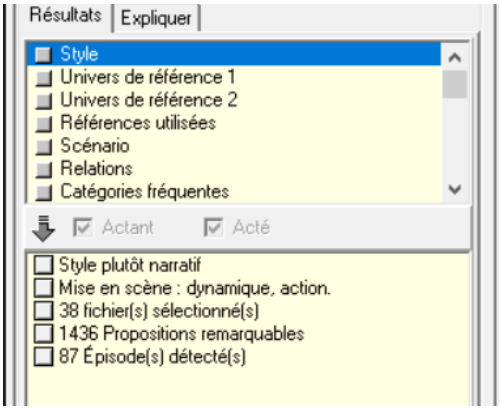



*Figure 1: Résultats des styles sur le logiciel Tropes*

Dans un second temps, ce sont les références qui seront analysées par le logiciel. Les références sont des objets, des concepts, des acteurs, des idées qui représentent le contexte. Elles sont regroupées selon des classes d'équivalents et selon trois niveaux : l'univers de référence 1 pour le premier niveau, l'univers de référence 2 comme second niveau et les références utilisées qui constitue le troisième et dernier niveau. Elles apparaissent par importance décroissante et sont accompagnées de leur nombre d'occurrences. Il est tout à fait possible de personnaliser la classification de ces références en ajoutant des éléments dans le scénario. Le scénario est une classification hiérarchique que le logiciel nous retourne en résultat. Il permet de personnaliser les dictionnaires du logiciels et structurer l'information comme nous le souhaitons. Pour créer ou modifier un scénario, il faudra utiliser l'outil scénario que l'on retrouve dans le menu [Outil]. Le logiciel ouvrira alors une nouvelle fenêtre destiné uniquement à ce scénario. Cet outil permet de construire notre propres classifications personnalisées. Celle-ci peut se faire manuellement à partir d'un scénario vide, automatiquement à partir du texte analysé ou nous pouvons ajouter un scénario déjà constitué au préalable. Dans ce dernier cas, il faudra sélectionner le fichier scénario qui doit avoir comme extension « .scn ». Dans un troisième temps, le logiciel est capable de nous retourner les relations qui peuvent exister entre deux termes et leur nombre d'occurrences. Les relations sont les mots qui sont retrouvés ensemble, dans le même ordre et dans la proposition du texte. Elles sont orientées suivant l'ordre d'apparition. Le dernier terme d'analyse à définir est celui des catégories. Le logiciel Tropes nous rapporte les catégories de mots du texte analysé. Il y a deux niveaux : les catégories fréquentes donc celles qui sont les plus utilisées, et toutes les catégories qui comportent les plus fréquentes et les moins fréquentes.




*Figure 3: Résultats de l'univers de référence 1 sur le logiciel Tropes*

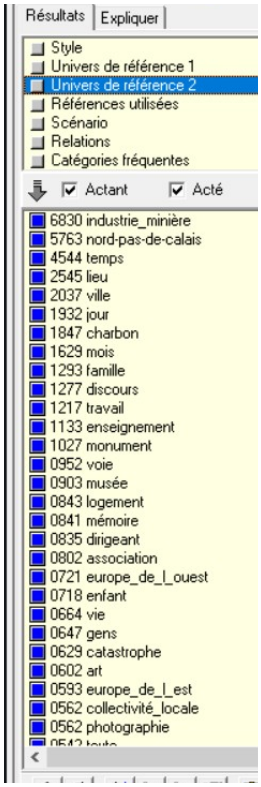



*Figure 2: Résultats de l'univers de référence 2 sur le logiciel Tropes*

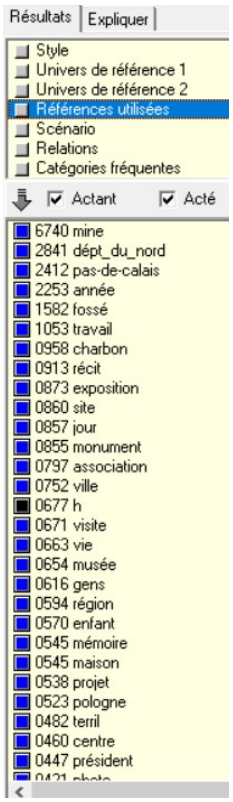



*Figure 4: Résultats des références utilisées sur le logiciel Tropes*

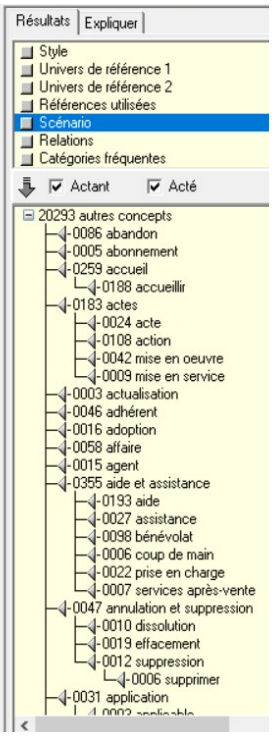



*Figure 6: Résultats du scénario sur le logiciel Tropes*




*Figure 5: Résultats des relations sur le logiciel Tropes*

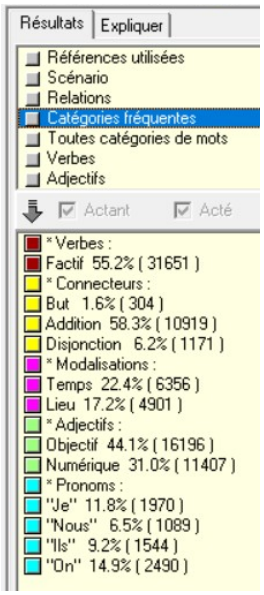



*Figure 7: Résultats des catégories fréquentes sur le logiciel Tropes*

Avec ces résultats sémantique plutôt complets, Tropes retourne aussi des graphes. Les graphes sont hypertextes, nous pouvons donc cliquer sur un élément pour voir le graphe qui lui est attribué. L'affichage est personnalisable, il suffit de déplacer son curseur pour faire varier le nombre de concepts affichés. Il y a cinq sortes de graphe :

- Le graphe-acteur qui est un graphe en fonction des acteurs sous forme donc de petites sphères nommées par un terme du texte et reliées entres elles si le graphe est une représentation des relations, par exemple. Sur ce graphe, ne sont représentées que les références qui présentent un nombre important de relation.
- Le graphe-aires qui est un graphe sous formes d'aires, c'est-à-dire que ces mêmes sphères sont disposés sur un système d'aires circulaires, comme notre système solaire, et leur surface varie en fonction du nombre de mots qu'elles contiennent. Ce graphe permet d'analyser l'environnement d'une référence ou d'une catégories. Les références sont orientés en fonction de leur place selon le terme central : à gauche il y aura tous les prédécesseurs du terme central et à droite tous les successeurs.
- Les graphes-étoilés, qui représentent un terme ou une relation dans le texte et tous les autres mots auxquels le terme central peut être relié. En d'autres termes, il affiche les relations entre les références ou entre une catégorie de mots et des références. Il affiche aussi la quantité de relation qui existe entre les références. Comme le graphe-aires, ce graphe analyse l'environnement d'une référence ou d'une catégorie et ce, de la même façon.
- Le graphe-répartition qui en fait, un histogramme qui montre la répartition d'une référence, d'une relation ou d'une catégorie de mot. Il est formé par la division du texte en secteurs contenant tous le même nombre de mot. Il représente la fréquence d'apparition du sujet à l'intérieur de chaque secteurs. Une barre représente un secteur et ils sont affichés par ordre chronologique. Ce graphe montre aussi une moyenne des barres grâce à une ligne en pointillés.
- Le graphe-épisodes affiche chaque rafale sous forme d'une ligne de pointillés qui indique son étendue et sa position. Les rafales sont affichées de haut en bas et de gauche à droite. Elles se succèdent selon leur ordre d'arrivée dans le texte. Ce graphe affiche aussi les épisodes sous forme de grands cadres pointillés.

Comme nous pouvons le constater, le logiciel permet une analyse très complète des données textuelles et diverses visualisations des résultats. C'est une des raisons pour laquelle nous nous sommes orientés vers celui. Mais ce n'est pas la seule.

#### 2.1.2.2 Pourquoi Tropes ?

Bien qu'il a été crée en 1994, et qu'il a donc 27 ans, Tropes reste un logiciel très intéressant pour effectuer une analyse de corpus d'articles de presse. Il a montré qu'il été capable d'analyser plus d'une trentaines de fichiers, plutôt conséquents, en même temps sans perdre de sa rapidité. Mais ce n'est pas non plus un des raisons principales du choix de Tropes.

La première raison de ce choix que nous pouvons expliquer est que Tropes est un logiciel français, créer par des français sur la base des travaux du psychologue français Rodolphe Ghiglione. Tout porte donc à croire qu'il a été créer dans le but premier de faire avancer les méthodes et les outils scientifiques de la recherche en France. Il a évidemment été développer dans plusieurs langues, comme l'anglais, l'espagnol, le portugais ou le brésilien, il n'est pas moins qu'il est une garantie pour analyser des textes écrits en langue française, ce qui n'est pas le cas pour certains outils développés dans d'autres pays.

La seconde raison est une question de proximité. En effet, malgré son âge et sa très petite notoriété, le laboratoire GERiiCO a une certaine proximité avec ce logiciel puisque certains de ses chercheurs l'utilisent toujours et il est enseigné aux étudiants des formations en Science de l'Information et du document, notamment dans le Master Veille et Communication de l'Information Stratégique (VeCIS) dans le cadre d'un cours sur l'information et l'aide à la décision où il est question de collecter et d'analyser l'information grâce à la pratique de l'analyse automatique avec le logiciel Tropes puis de la construction de scénario et

d'analyse de corpus. Le but de ce cours étant d'apprendre aux étudiants à utiliser un outil d'analyse de l'information.

Nous arrivons donc à la dernière raison qui constitue ce choix qui est la possibilité de construire un scénario. Comme expliqué dans la partie précédente, la logiciel permet d'utiliser une classification hiérarchique des références pour l'analyse des textes. Nous avons le choix de créer notre propre classification directement avec le logiciel, en choisir un établi au préalable ou de laisser le logiciel le créer lui même à partir des termes des textes. C'est une fonctionnalité qui représente un atout par rapport à d'autres logiciels comme Iramuteq qui propose une unique méthode de classification hiérarchique, pour ne prendre que cet exemple.

Finalement, bien des outils d'analyse sémantique étaient disponibles pour nous aider dans l'analyse de corpus de presse et la création d'une grille d'analyse mais un seul validait tous les critères de sélection pour cette tâche. Sachant aussi que comme il est étudié dans un des Master du département des Sciences de l'Information et du Document, il constituait une opportunité, pour le stagiaire choisi, d'avancer dans l'apprentissage de l'utilisation de cet outil, qu'il aura l'occasion de rencontrer quelques mois plus tard, ainsi que de connaître la bonne méthodologie à utiliser.

## 2.2 Méthodologie utilisée

Il existe une méthodologie générale de l'analyse de contenu et elle consiste en deux étapes selon André Désiré Robert et Annick Bouillaguet dans le chapitre II de leur ouvrage « L'analyse de contenu », publié en 2007. La première étape consiste à une travail en amont de l'analyse qui est de définir précisément le sujet à traiter et ainsi d'élaborer une problématique autour de ce sujet. Le but de la problématique est d'essayer de traiter sous un nouvel angle l'objet de recherche et d'apporter de nouvelles réponses. La seconde étape de cette méthodologie générale est donc l'analyse de contenu en elle-même qui comporte à son tour plusieurs phases. La première phase de la seconde étape est la pré-analyse qui est déterminer sur quel support établir l'analyse. La seconde phase est celle de la catégorisation. Elle a comme objectif de traiter une première fois les documents à analyser en catégorisant les éléments les plus pertinents du corpus. Ces catégories doivent être pertinentes, exhaustives, exclusives et objectives. La troisième phase est le codage et le comptage des unités et la dernière phase mène l'interprétation des résultats.

Afin de mieux comprendre ce qu'est la méthodologie à adopter dans le cadre d'une analyse de corpus de presse, nous allons définir la méthodologie utilisée pour le cas de l'analyse de corpus de presse pour le projet ANR MémoMines. Nous rentrerons ensuite plus en détails dans cette méthodologie et présenter quels sont les critères qui permettront d'analyser ce corpus de presse, puis terminer sur le sujet de la grille d'analyse qui est la finalité de cette analyse de contenu et, évidemment, du stage.

A mon arrivée en tant que stagiaire sur le projet ANR MémoMines, une partie du corpus à analyser avait déjà été sélectionné et récupéré via le serveur Europresse qui permet aux professionnels de l'information et de la communication de consulter un répertoire régulièrement mis à jour d'articles de presse. Ce serveur est donc un répertoire des sources en ligne de plus de 55 000 sources reconnus. Il a été crée par la société Cision qui est leader mondial des technologies de l'information et de la veille. En plus de la récupération du corpus qui avait déjà commencé, des fichiers indiquant des listes de lieux et d'acteurs, ayant joués un rôle pendant et après les siècles d'exploitation du charbon, avaient aussi déjà été rédigés. Le but de cette analyse était de confirmer ces listes établies, et de les modifier si un nouvel acteur ou un nouveau lieu était retrouvé dans le corpus d'articles de presse. Aussi, un thesaurus a été créée dans le but d'avoir une représentation terminologique du patrimoine minier.

La première tâche qui m'a été attribué dans mon stage a été de prendre connaissance du projet et ses enjeux. Une documentation constituée de documents scientifiques m'a été fourni afin de connaître le contexte dans lequel ma mission de stage s'inscrit. Ensuite, il a fallut compléter le corpus. En effet, les articles de la seconde partie de l'année 2020 et des premiers mois de l'année 2021 n'avait pas encore été collecté. Pour ce faire, j'ai du suivre la méthode de construction du corpus qui m'a été indiqué. Cette méthode consiste à interroger le répertoire Europresse selon trois requêtes. Pourquoi trois requêtes ? Le serveur Europresse ne donne pas la possibilité d'exporter un nombre illimité d'articles c'est pourquoi il a fallut scinder les exportations en trois lots, selon la longueur des articles, c'est-à-dire : les articles courts comprenant entre 100 et 297 mots, les articles moyens entre 304 et 695 mots puis les articles longs. Dans ces premières données, on constate que les intervalles ne sont pas très précises car Europresse ne donne pas de définition stricte de la longueur des articles. Voici donc les requêtes permettant l'exportation des articles :

1. **Articles courts** : TIT_HEAD= (charbon | mines | mine | minier | mineur | fosse) & LG= court & LEAD= ((charbon | mines | mine | minier | mineur | fosse)! ("le fossé" | "au fossé")) ! TEXT= ("école des mines" | antipersonnel | "mine de rien" | "grise mine" | noeux | douchy | bully | auchy | halluin | chine | marles | enquin | carmaux | nomades | monoxyde)
2. **Articles moyens** : TIT_HEAD= (charbon | mines | mine | minier | mineur | fosse) & LG=moyen & LEAD= ((charbon | mines | mine | minier | mineur | fosse)! ("le fossé" | "au fossé"))! TEXT= ("école des mines" | antipersonnel | "mine de rien" | "grise mine" | noeux | douchy | bully | auchy | halluin | chine | marles | enquin | carmaux | nomades | monoxyde)
3. **Articles longs** : TIT_HEAD= (charbon | mines | mine | minier | mineur | fosse) & LG=long & LEAD= ((charbon | mines | mine | minier | mineur | fosse)! ("le fossé" | "au fossé")) ! TEXT= ("école des mines" | antipersonnel | "mine de rien" | "grise mine" | noeux | douchy | bully | auchy | halluin | chine | marles | enquin | carmaux | nomades | monoxyde)

Les mots compris dans les requêtes sont les termes qui devraient figurer dans les articles de presse afin qu'ils soient pertinents. Elles sont larges et ont été créées volontairement dans ce sens. Elles entraînent donc du bruit qu'il faudra réduire en sélectionnant manuellement les articles en effectuant une lecture des titres et des premières lignes des articles. Aussi, il faudra ajouter des filtres dans la recherche, notamment pour définir la source qui est ici La Voix du Nord et faire une recherche selon les années afin d'avoir un document pour une année d'articles.

Suite à la récupération du corpus, tout d'abord sous le format PDF, il a fallut nettoyer les documents. Europresse propose de télécharger les titres sous le format PDF ou le format HTML. Le format PDF occasionnent du bruit, notamment une page de garde avec des notes de la Société Cision, une ou plusieurs pages de sommaires en fonction du nombre d'articles sélectionnés pour une année, certaines métadonnées qui ne sont pas utiles à l'analyse notamment des certificats d'utilisations et autres métadonnées juridiques. C'est là le premier problème que nous avons rencontré. Tout d'abord j'avais personnellement un problème à faire entrer les fichiers PDF sur le logiciel Tropes. Il donc fallut trouver un convertisseurs qui pouvaient supporter une bonne centaine de fichiers à convertir en même temps et convertir correctement. Nous n'avons pas trouvé de convertisseurs qui convenait à nos attentes. Aussi, le nettoyage d'un corpus en format PDF nécessitait de nombreuses manipulations qui nous auraient fait perdre beaucoup de temps car elles n'étaient pas encore totalement connues. Un choix s'est donc posé à nous, soit continuer à pratiquer plusieurs manipulations sur du PDF, soit récupérer tout le corpus en HTML et pratiquer des transformations plus précises et plus rapides sur les fichiers en utilisant le langage XSLT. Nous avons donc opté pour la seconde option. Une fois le corpus récupérer totalement en HTML, nous avons transformés ces fichiers en ne gardant que les éléments pertinents pour l'analyse et dans les formats souhaités. Une première transformation du corpus a été effectuée qui est une transformation des documents HTML en format texte brut ne comprenant uniquement le titre des articles et leur contenu pur.

xsl:stylesheet

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
3      xmlns:xs="http://www.w3.org/2001/XMLSchema"
4      exclude-result-prefixes="xs"
5      version="2.0">
6      <xsl:output method="text" encoding="iso-8859-15"/>
7      <xsl:template match="body/article" name="regle1">
8          ------------------------------------
9          <xsl:for-each select="header/div/p">
10             <xsl:if test="@class='titreArticleVisu rdp__articletitle'">
11                 <xsl:value-of select="."/>
12             </xsl:if>
13         </xsl:for-each>
14         >&#x09;
15         <xsl:for-each select="section/div/div/p">
16             <xsl:value-of select="."/>
17         </xsl:for-each>
18     </xsl:template>
19 </xsl:stylesheet>
```

*Figure 8: Script de la transformation en texte brut*

Résultats

```
1   html,body{font-family:Arial,'tahoma';font-size:11px;background-color:#fff;margin:5p
2       ------------------------------------
3       Rafraîchissement autour du Mémorial du  mineur... avant son déménagement ?
4       >
5       par Anne-Claire Guilain bethune@lavoixdunord.fr Bruay-La-Buissière. Vous avez s
6       ------------------------------------
7       À l'époque où les pompiers des  mines intervenaient en torpédo
8       >
9       Auberchicourt. Que de visages, aujourd'hui disparus. Le service d'incendie comp
10      ------------------------------------
11      Le bassin  minier et ses artistes inspirent le collectif  Mines de rien
12      >
13      par christophe lecouteux lens@lavoixdunord.fr Hénin-Beaumont. L'un est spéciali
14      ------------------------------------
15      Comment le plateau de « Germinal » est arrivé au Centre historique  minier
16      >
17      par Julien Gilman douai@lavoixdunord.fr lewarde. Le Covid ne réserve pas que de
18      ------------------------------------
19      Accusé de délaisser les affiliés au régime  minier, Filieris répond
20      >
21      Par Élise Forestier lens@lavoixdunord.fr Liévin. Début février, Laurent Duporge,
22      ------------------------------------
23      Grand défenseur du bassin  minier, Gilbert Rolos, ancien maire, est décédé
24      >
25      Par Céline Debette lens@lavoixdunord.fr Sallaumines. Le 30 juin 2012, à Saint-P
26      ------------------------------------
27      Terrils avec vue sur site  minier UNESCO cherchent vaches pour pâturer
28      >
29      Par Anna Morello lens@lavoixdunord.fr Oignies. Cherche six à huit occupants pou
30      ------------------------------------
```

XSL - xml - transfo_txt.xsl (transfo_txt1)

*Figure 9: Résultats de la transformation en texte brut*

Deux
autres transformations ont été effectuées. L'une en gardant le format HTML et l'autre en basculant sur le format XML. Dans ces deux nouvelles versions, le corpus contient le titre et le contenu des articles, mais aussi le nom de la source et la ville, la date de publication et le numéro de l'article.

```
    version="2.0">
    <xsl:output method="html" encoding="UTF-8"/>
    <xsl:template match="body" name="regle1">
        <html>
            <head>
                <title>Resultats de la transformation</title>
            </head>
            <body>
                <header>
                </header>
                <section>
                    <xsl:for-each select="article">
                        <article>
                            <xsl:for-each select="header/div/p">
                                <xsl:if test="@class='titreArticleVisu rdp__articletitle'">
                                    <h1><xsl:value-of select="."/></h1>
                                </xsl:if>
                            </xsl:for-each>
                            <xsl:for-each select="header/div/span">
                                <xsl:if test="@class='DocPublicationName'">
                                    <h2><xsl:value-of select="."/></h2>
                                </xsl:if>
                                <xsl:if test="@class='DocHeader'">
                                    <h3><xsl:value-of select="."/></h3>
                                </xsl:if>
                            </xsl:for-each>
                            <xsl:for-each select="section/div/div/p">
                                <p><xsl:value-of select="."/></p>
                            </xsl:for-each>
                            <xsl:for-each select="footer/div/div">
                                <xsl:if test="@class='publiC-lblNodoc'">
                                    <p><xsl:value-of select="."/></p>
                                </xsl:if>
                            </xsl:for-each>
                        </article>
                    </xsl:for-each>
                </section>
            </body>
        </html>
    </xsl:template>
</xsl:stylesheet>
```

Texte Grille Auteur

*Figure 10: Script de la transformation en HTML*

Résultats

```
    <meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
    <title>Resultats de la transformation</title>
</head>
<body>
    <header></header>
    <section>
        <article>
            <h1>L'odyssée du  charbon</h1>
            <h2> La Voix du Nord  DOUAI </h2>
            <h3>  dimanche 19 septembre 2004 182 mots </h3>
            <p>Au Centre historique minier de Lewarde, les Journées du Patrimoine, sont consacrées
               cette année au thème des sciences techniques. Ce dimanche, les visiteurs pourront
               accéder gratuitement aux différentes expositions permanentes et temporaires du musée,
               ainsi qu'à deux animations,  Histoire de la fosse Delloye  et  L'Odyssée de la vie
               sur Terre  présentées au cours de cette journée par les médiateurs culturels du Centre.
               </p>
            <p> L'animation  Histoire de la fosse Delloye  entraînera le public dans une promenade
               guidée du carreau où seront notamment évoquées les caractéristiques architecturales
               et les infrastructures d'une fosse des années 1930...Quant à  L'odyssée de la vie
               sur Terre , elle permettra de découvrir les différentes phases de l'évolution de la
               vie sur Terre en parcourant les 455 mètres de l'échelle des temps géologiques géante.
               Les visiteurs pourront apprécier en particulier certaines étapes importantes de cette
               odyssée, comme l'apparition de la vie il y a 3,8milliards d'années, puis les premiers
               vertébrés... pour terminer avec l'hominisation. Par ailleurs, la visite des galeries
               guidée par un ancien mineur sera proposée à demi-tarif soit 5,30Euro(s)E pour les
               adultes et 3,70Euro(s)E pour les enfants, de 9 heures à 19h30 avec la fermeture de
               la billetterie à 17h30. Enfin, les deux expositions temporaires présentées lors de
               ces Journées du Patrimoine sont liées à l'immigration dans le Bassin minier du Nord-Pas-de-Calais:
               Ahmed, Wladislaw, Dario... tous «gueules noires»</p>
            <p>et  tout un spectacle: collection d'affiches polonaises d'Henri Juskowiak.</p>
            <n>news·20040919·VN·20040919571</n>
```

*Figure 11: Résultats de la transformation en HTML*




*Figure 12: Script de la transformation en XML*

Après avoir nettoyé le corpus et extrait tous les éléments jugés pertinents pour l'analyse, nous avons entrer le nouveau corpus en texte brut sur Tropes afin de faire une première analyse des résultats retournés par le logiciel. Un nouveau problème est survenu, celui du scénario renvoyé par Tropes. Le scénario proposé par Tropes ne correspondait pas à ce que nous souhaitions retirer de l'analyse. En effet, le premier scénario que nous retourne Tropes est un scénario qu'il a lui-même créer grâce à ses dictionnaires et en fonction des éléments qu'il retrouve dans le corpus. C'est une fonctionnalité très utile lorsque l'on cherche à trouver des éléments inconnus mais dans notre cas, rappelons-le, nous cherchons à confirmer les listes préétablies par les collègues avant le début du stage. Il a donc fallut créer des nouveaux scénarios à partir de scénarios vides pour les lieux et les acteurs, et inscrire directement des codes de structuration dans le fichier du thesaurus. La création de nouveaux scénarios à partir d'un fichier vide nécessite une réflexion sur la structuration. Nous avons choisi pour le scénario des lieux de classer les villes de la région Hauts-de-France en fonction de l'arrondissement dans lequel elles font parties, et du département dans lequel les arrondissements font partis.

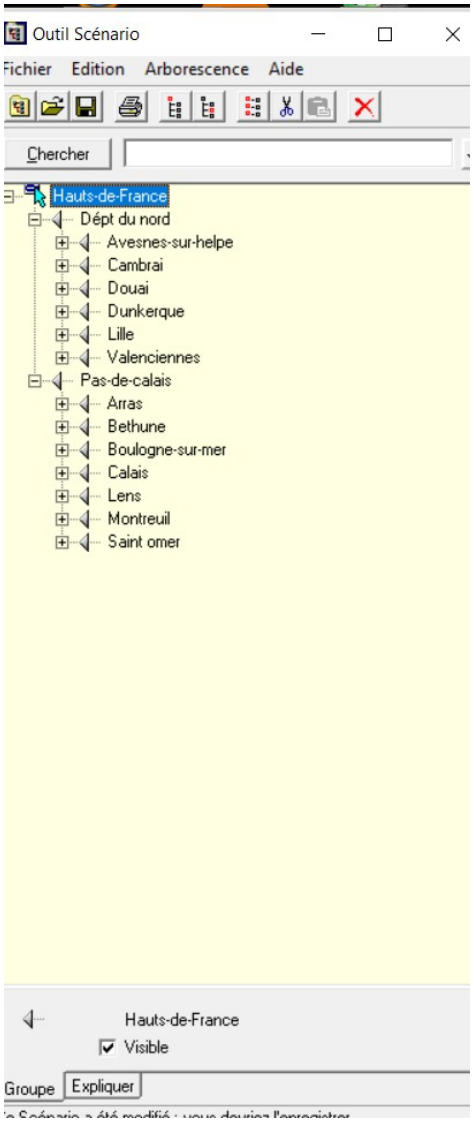



*Figure 13: Scénario des lieux 1*

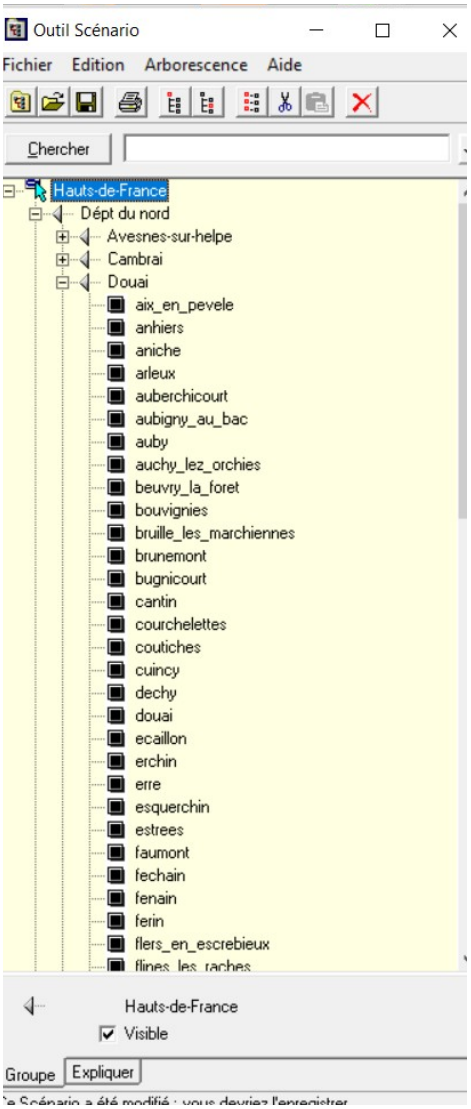



*Figure 14: Scénario des lieux avec détail sur les villes*

Pour les acteurs, nous avons choisi de structurer le scénario en fonction de si ils sont des associations, des musées ou des sites miniers et pour chacun, en fonction de leur région et de leur département.

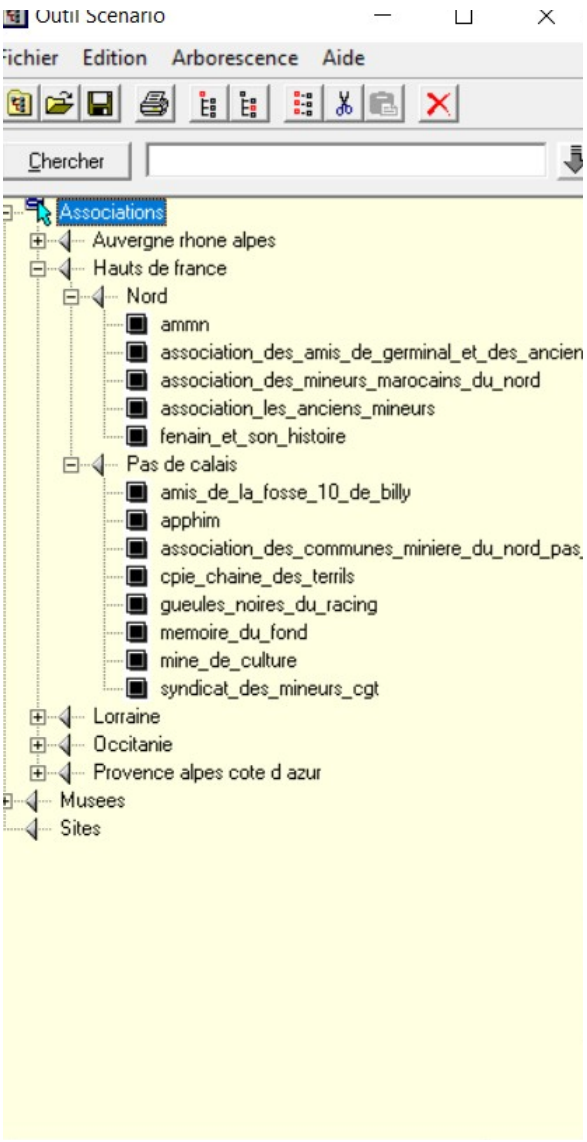



*Figure 15: Scénario des acteurs*

Le thesaurus a été recopié entièrement sur un fichier.scn. Il était possible aussi d'ouvrir le fichier du thesaurus sur un bloc-note et y ajouter des codes. Les codes à ajouter dans le fichier du thesaurus correspondent à position dans le thesaurus de l'élément codé. Il existe des codes en fonction des niveaux et en fonction de leur nature, soit si un élément est un groupe qui contient d'autres éléments, ou si un élément est un terme uniquement.

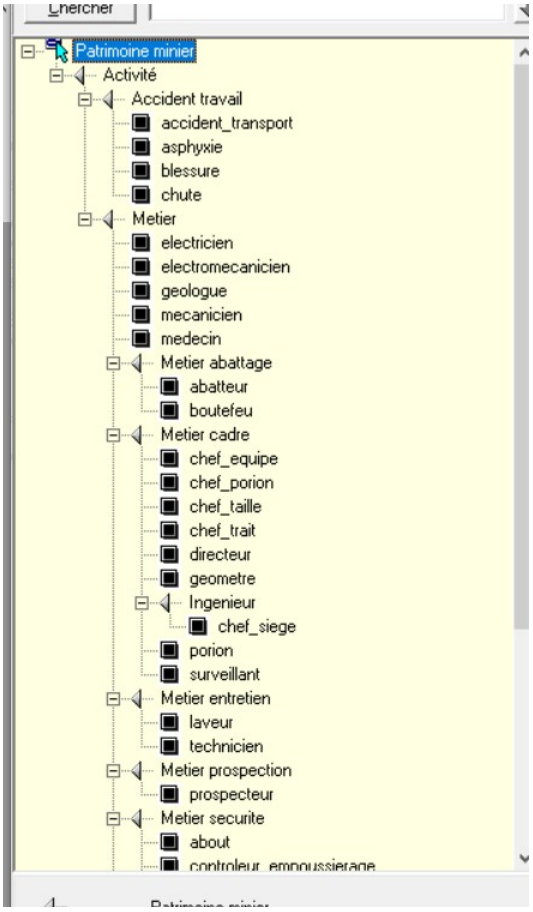



*Figure 16: Scénario du thesaurus du patrimoine minier "ThesoMines"*

Les scénarios se définissent ici comme des grilles d'analyse du contenu dans le sens où nous avons établit une liste des grandes catégories et des unités correspondantes, qui peut aussi bien servir à un corpus d'articles de presse qu'à d'autres corpus textuels. Ces scénarios, qu'ils soient utilisés seuls ou bien combinés, permettent à l'analyste d'évaluer le corpus de la façon la plus objective possible, bien qu'ils ne suppriment pas totalement la subjectivité. Cela dit, bien que les scénarios permettent de confirmer les listes d'acteurs et de lieux déjà trouvés dans des analyses d'autres corpus, ils ne permettent pas d'en définir de nouveaux. Ils marquent effectivement les unités déjà déterminés au cours du projet mais d'autres unités peuvent encore se cacher dans le corpus.

Pour conclure cette partie sur la méthodologie, nous justifierons le choix de l'utilisation de l'outil Tropes pour l'analyse sémantique par le fait que le logiciel nous donnait accès aussi bien aux résultats que le logiciel obtenait de lui-même qu'à la possibilité de créer nos propres catégories mais surtout nos propres grilles d'analyse. Bien que le logiciel est un dictionnaire plutôt large, les propositions de références ne convenaient pas totalement aux références que nous possédions au préalable, et ne permettaient pas de produire une analyse qui tendait vers ce que nous cherchions. C'est un logiciel très utile qu'il faut continuer d'améliorer notamment peut-être au niveau de la création des scénarios, dans le sens où cela serait plus pratique de pouvoir intégrer un scénario à partir d'un fichier autre qu'avec l'extension « .scn », qu'il est délicat de créer en dehors du logiciel. Cependant, c'est un outil qui favorise énormément l'objectivité dans les analyses, étant donné que l'humain n'est pas spécialement obligé de faire de nombreuses manipulations pour avoir des résultats satisfaisants.

# 3 Structuration de l'information

La structuration de l'information représente toutes les techniques de production, de stockage et d'accès au contenu. Elle fait une certaine différence entre les bases de données et les corpus documentaire dans le sens où les techniques vont différer selon si l'on souhaite retrouver le contenu de chaque champ d'une base de données ou si nous souhaitons une recherche d'information en texte intégral. Structurer un document permet d'inscrire le contenu dans une forme afin de le retrouver grâce à des index spécifiques par exemple, ou de le publier sous des formes diverses. Une structuration logique permet de définir des éléments hiérarchisés et d'assigner des attributs pour chaque éléments.

Les objectifs de la structuration de l'information sont de permettre un meilleur traitement, ici automatique, des données. Pour rappel, une donnée est une information qui n'a pas été transformée pour être traitée dans le but de leur donner un sens, d'être communiquée. Une structuration des données peut être différente en fonction du traitement informatique que l'on souhaite lui approprier, et cela faciliterait le travail d'application information et diminuer le nombre d'erreurs de traitement.

Dans notre cas, nous parlerons essentiellement de structuration de données numériques puisque les données à traiter sont celles du corpus d'articles de presse issus de La Voix du Nord et récupérés sous format HTML via le serveur Europresse.

Dans le cadre du stage, la structuration de l'information va être utilisée afin de produire une base de connaissance dans le but de valoriser le patrimoine minier. Les données étant diffuses dans les contenus de la presse, la première tâche a donc été de collecter les informations, pour ensuite les structurer avant de les analyser. Il y a donc des enjeux de préparation du corpus et d'identification des éléments pertinents. Plus précisément, et comme expliqué de manière plus abstraite dans la méthodologie, il a fallut transformer et structurer de façon automatique le corpus d'articles de presse afin de permettre au logiciel Tropes un meilleur traitement des documents pour l'analyse de leur contenu mais aussi dans le but de réutiliser ces documents dans le projet sans pour autant être gêner par le bruit des métadonnées non pertinentes à quelconques analyse.

Dans cette partie, nous essaierons de savoir comment s'intègre la structuration de l'information dans une analyse de contenu ? Comment sont redistribués les rôles du balisage et des langages informatiques dans la structuration de l'information puis dans une analyse de corpus d'articles de presse ?

Afin de répondre à cette problématique, nous allons tout d'abord présenter quelques définitions des notions de base dans la structuration des informations afin de mieux cadrer le sujet et nous focaliser sur les notions que nous rencontrerons au fur et à mesure des explications. Nous démontrerons ensuite le rôle du balisage dans la structuration de l'information mais aussi son rôle dans le cadre de l'analyse de corpus de presse dans le projet ANR MémoMines. Et nous terminerons par aborder les langages informatiques qui nous ont été utile pour le traitement numérique des articles de presse, notamment les langages XSLT, HTML et XML.

## 3.1 Définitions

Pour commencer, nous allons définir très rapidement les notions d'information, notamment en tant que signal, de systèmes d'information, de données et de document.

L'information est « *une connaissance inscrite (enregistrée) sous forme écrite (imprimée ou numérisée), orale ou audiovisuelle sur un support spatio-temporel. L'information comporte un élément de sens. C'est une signification transmise à un être conscient par le moyen d'un message inscrit sur un support : imprimé, signal électrique, onde sonore, etc.* » (Le Coadic, 2004). Selon Shannon, l'information est un flux physique circulant entre un émetteur et un récepteur lors d'un processus de communication, comme

l'indique le Modèle de Shannon et Weaver[3]. Dans ce modèle, la source d'information énonce un message que l'émetteur va encoder et transformer en signal qui va être acheminé par le canal, puis décodé par le récepteur, qui reconstitue un message à partir de ce signal et le transmet au destinataire.

L'information est une sorte de séries de codes, comme par exemple le fait d'afficher un titre sous une grande police,de surligner des éléments importants dans le contenu, ou de créer une certaine hiérarchie des titres, qui permettent de définir la structure de l'information ou du document.

Pour parler de la notion de système d'information, nous allons surtout nous intéresser à son enjeu informatique. Un système d'information est « *un ensemble organisé de ressources qui permet de collecter, stocker, traiter et distribuer de l'information...* » (Wkipédia). Dans notre cas, les systèmes d'information vont être toutes les solutions informatiques qui vont permettre de collecter, stocker, traiter et communiquer les informations. Ces systèmes vont faire appel à des langages informatiques en fonction de leur utilité, par exemple, pour permettre une meilleure gestion des informations, nous nous orienterons vers du langage Java ou SQL. Dans le stage, l'enjeu est de baliser l'information pour un traitement automatique, nous utilisons donc les langages de balisage comme HTML ou encore XML.

Un document est l'ensemble formé par une information et son support. Il est fabriqué dans le but d'expliquer, de décrire et peut être utilisé comme une preuve. Sa forme numérique est celle qui nous intéresse le plus ici. Le document numérique est celui qui est utilisé dans ce stage. Sous cette forme, il permet une séparation entre les métadonnées, soit toutes les informations relatives au document, que ce soit sa nature, son auteur, sa date de création, etc. et le contenu, soit les informations destinées à être communiquée par ce document, comme par exemple des informations textuelles, des images, des tableaux.

## 3.2 Balisage et langages de structuration de l'information

Avant d'expliquer les termes de balisage et de langages de structuration de l'information, il est important de préciser que notre travail s'appuie sur la modélisation de l'information. La modélisation est « *une technique d'ingénierie visant à comprendre un système, déjà existant ou à créer. Elle permet de « visualiser » […] un système tel qu'il est, ou tel que nous voudrions qu'il soit ; d'en préciser la structure ou le comportement suivant des points de vue qui éclairent la réalité de différentes façons, et ceci indépendamment d'un langage de programmation* » (Dalbin, 2003). La modélisation permet donc de structurer les idées et simplifier la réalité dont la représentation est abstraite. En d'autres termes, le but est de construire un système pour le documenter. En prenant l'exemple de notre mission, il a fallut modéliser et donc de structurer le corpus d'articles de presse en balisant totalement les fichiers afin de repérer très rapidement les titres des articles, le nom de leur source, leur numéro, leur date de publication et évidemment leur contenu. Il est plus aisé pour l'humain de se concentrer sur une zone du document à la fois plutôt que sur un ensemble de données non structurées limitant la perception ne serait-ce que d'un unique élément de l'information. Pour faire une comparaison avec une situation du quotidien, l'humain a du mal à retrouver l'objet qu'il cherche dans une maison qui n'est pas rangé alors que dans une maison bien rangé, il visualise déjà la pièce dans lequel l'objet peut se trouver éliminant ainsi toutes les autres pièces de la maison.

Dans un langage informatique, une balise permet de repérer une position dans un processus de structuration de l'information dans un document. Elle marque l'emplacement de cette information par rapport au flux d'information que propose un document. Si plusieurs information doivent être marquées, il faudra utiliser plusieurs types de balise. Dans ce cas, la balise n'est plus seulement un marquage mais devient un élément d'information aussi important que le contenu du document.

Dans ce même contexte, le balisage permet de définir une zone dans le document. Cela permet de repérer rapidement la partie du document qui nous intéresse étant donné qu'elle possède une caractéristique

3 Cf Modèle de Shannon et Weaver en Annexe

particulière qui la différencie du reste du contenu. Sans parler de langage de balisage, le moyen le plus abordable de baliser une information dans un document est l'application de style dans un texte, que ce soit la mise en gras, en italique, entre parenthèses ou entre guillemets. Plus techniquement, le balisage est le fait de mettre entre deux balises indiquant le début et la fin de la zone à marquer, une information. On parle ici de balises ouvrante et de balises fermantes. Dans certains langages, une balise ouvrante doit absolument correspondre à une balise fermante.

Dans le cadre du stage, nous avons eu l'occasion de pratiquer le balisage par des langages de structuration de l'information, notamment grâce aux langages informatiques sur lesquels nous reviendront plus tard. Nous avons retrouver cet aspect de la structuration de l'information notamment au moment de la transformation du corpus. Au vu des problèmes rencontrer lors du nettoyage et de l'analyse sémantique de Tropes, il a fallut trouver une solution afin de mener ces tâches le plus rapidement possible et de manière automatique.

Comme expliqué plus généralement dans la partie méthodologie de ce mémoire, nous avons dû procéder à un nettoyage du corpus afin de permettre un meilleur traitement par le logiciel Tropes, et limiter donc les erreurs, mais aussi pour que ce corpus puisse être réutiliser à l'avenir sans que les futurs réutilisateurs soient gêné par le bruit que pouvait comporter la version originale téléchargée en HTML sur le serveur Europresse. Le corpus de la version originale en HTML a donc été créer en trois autres nouvelles versions : la première en texte brut, pour une utilisation à court terme dans le logiciel Tropes, comportant uniquement le titre des articles et leur contenu ; la seconde en HTML, pour une utilisation à long terme dans d'autres analyses, incluant en plus des titres des articles et leur contenu, leur date de publication, leur source et leur numéro ; pour finir la troisième version en XML, pour la même utilisant et comportant les mêmes éléments que la version en HTML. Afin de procéder à ces transformation, il a fallut apprendre à utiliser le langage XSLT. Le langage XSLT a été conçu pour transformer des documents XML en document d'autres formats. Il permet de créer des règles de transformation[4] sur un document donné en entrée pour générer en sortie un nouveau document dans le format que l'on souhaite. Un document XSLT est en fait un document XML que l'on peut aussi appeler comme feuille de style XSLT. Une feuille de style XSLT possède une structure de base comprenant un prologue et un élément racine. Les autres éléments qui seront ajoutés sur cette feuille devront commencer par xsl :, comme par exemple, l'élément xsl:output, qui permet de préciser les caractéristiques de sortie du document à créer, ou l'élément xsl :template, qui définit le nom des règles à appliquer et sur quelle partie du document d'entrée la règle doit être appliquée, que l'on peut voir ci-dessous :

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
3      xmlns:xs="http://www.w3.org/2001/XMLSchema"
4      exclude-result-prefixes="xs"
5      version="2.0">
6      <xsl:output method="html" encoding="UTF-8"/>
7      <xsl:template match="body" name="regle1">
```

*Figure 17: Exemple de structure de base d'une feuille de style XSLT*

Aussi, afin de générer un document, il est important de connaître le langage utilisé pour tel ou tel format et de respecter les styles attribués à ce dernier. Par exemple, si le document que l'on souhaite générer sortira sous format HTML, il faudra structurer le document comme un document HTML :

4 *cf schéma d'une transformation par XSLT du cours Cours3-XSLT.pptx*

```
<xsl:template match="body" name="regle1">
    <html>
        <head>
            <title>Resultats de la transformation</title>
        </head>
        <body>
            <header>
            </header>
            <section>
                <xsl:for-each select="article">
                    <article>
                        <xsl:for-each select="header/div/p">
                            <xsl:if test="@class='titreArticleVisu rdp__articletitle'">
                                <h1><xsl:value-of select="."/></h1>
                            </xsl:if>
                        </xsl:for-each>
                        <xsl:for-each select="header/div/span">
                            <xsl:if test="@class='DocPublicationName'">
                                <h2><xsl:value-of select="."/></h2>
                            </xsl:if>
                            <xsl:if test="@class='DocHeader'">
                                <h3><xsl:value-of select="."/></h3>
                            </xsl:if>
                        </xsl:for-each>
                        <xsl:for-each select="section/div/div/p">
                            <p><xsl:value-of select="."/></p>
                        </xsl:for-each>
                        <xsl:for-each select="footer/div/div">
                            <xsl:if test="@class='publiC-lblNodoc'">
                                <p><xsl:value-of select="."/></p>
                            </xsl:if>
                        </xsl:for-each>
                    </article>
                </xsl:for-each>
            </section>
        </body>
    </html>
</xsl:template>
</xsl:stylesheet>
```

*Figure 18: Exemple d'une feuille de style XSLT structurée pour une sortie HTML*

Sur cette image, on distingue bien la structure attribuée au format HTML avec ses balises : <html>, <head>, </head>, <body>, <section>, … On aperçoit aussi au sein de cette structure que des éléments de langage XSLT apparaissent. La balise <section> va concerner tous les articles se trouvant dans le document HTML du corpus orignal. Pour chaque article, nous mettrons dans la balise <article> du document de sortie, les titres (‘*titreArticleVisu rdp_articletitle*’), le nom du document de publication (‘*DocPublicationName*’), etc.

On retrouve ces mêmes éléments dans la feuille de style XSLT pour la transformation du corpus HTML original en XML. La structure attribuée au format XML est respectée et les éléments du langage XSLT sont incorporés dans cette structure sans pour autant la déformer :

xsl:stylesheet

```
        <resultats>
            <article>
                <titre>
                    <xsl:for-each select="header/div/p">
                        <xsl:if test="@class='titreArticleVisu rdp__articletitle'">
                            <xsl:value-of select="."/>
                        </xsl:if>
                    </xsl:for-each>
                </titre>
                <source>
                    <xsl:for-each select="header/div/span">
                        <nom_lieu>
                            <xsl:if test="@class='DocPublicationName'">
                                <xsl:value-of select="."/>
                            </xsl:if>
                        </nom_lieu>
                        <date>
                            <xsl:if test="@class='DocHeader'">
                                <xsl:value-of select="."/>
                            </xsl:if>
                        </date>
                    </xsl:for-each>
                </source>
                <contenu>
                    <xsl:for-each select="section/div/div/p">
                        <xsl:value-of select="."/>
                    </xsl:for-each>
                </contenu>
                <numero>
                    <xsl:for-each select="footer/div/div">
                        <xsl:if test="@class='publiC-lblNodoc'">
                            <xsl:value-of select="."/>
                        </xsl:if>
                    </xsl:for-each>
                </numero>
            </article>
        </resultats>
    </xsl:template>
</xsl:stylesheet>
```

*Figure 19: Exemple d'une feuille de style XSLT structurée pour une sortie XML*

# 4 Conclusion et perspectives

Pour conclure ce mémoire, l'analyse de corpus de presse s'intègre dans un projet de valorisation du patrimoine minier dans le sens où, si elle est effectuée seule, sans analyse d'autres données au préalable, elle peut faire ressortir des éléments pertinents dans un contenu en rapport avec le domaine de la mine. Elle peut extraire les acteurs du domaine de la mine mais aussi, les lieux, les évènements ou tout autre objet qui apparaîtrait répétitivement dans le corpus. Alliée à d'autres analyses et autres recherches elle permet aussi de confirmer ou de démentir leurs résultats.

Pour répondre à la question des critères de sélection pour la collecte d'articles de presse, nous avons utilisé le serveur Europresse qui nous permettait de naviguer à travers les archives de la presse papier et numérique. Nous avons choisi de tenir compte des articles provenant uniquement publiés par La Voix du Nord sur le domaine de la mine entre les années 2004 et début d'année 2021. Pour trouver les articles sur Europresse, nous avons utilisés des requêtes permettant de trouver des articles, tout d'abord en fonction de leur longueur (court, moyen, long) et contenant des termes tels que « mines », « charbon », « fosse » pour les titres ; et des termes tels que « halluin », « douchy » ou « monoxyde » pour le contenu.

La structuration de l'information a été une étape importante dans cette analyse puisque pour nettoyer le corpus, il a fallut utiliser des langages de structuration et de balisage pour arriver à générer des fichiers capables d'être analysés par un logiciel sans faire ressortir d'erreurs mais aussi que l'on puisse réutiliser à l'avenir dans ce même projet ou pour d'autres recherches.

Il y avait plusieurs outils capables d'automatiser une analyse sémantique, mais Tropes a été le logiciel le plus intéressant pour ce cas de figure, notamment, par sa proximité avec les chercheurs du laboratoire GERiiCO mais aussi leurs collègues qui travaillent sur leur mise à jour. Cela nous a permis de nous former très rapidement sur l'outil et de ne pas tarder pour faire ressortir des résultats, étant donné que le la mission de stage a une durée de deux mois, le nettoyage du corpus ayant pris beaucoup de temps. De plus, le logiciel nous permettait de créer des scénarios selon nos besoins et nos recherches. Nous avons donc pu élaborer trois scénarios :

- un premier scénario arborant les villes du Hauts-de-France selon l'arrondissement, le département, et la région auxquels elles appartiennent ;
- un second scénario pour les acteurs, selon une liste prédéfinie. Il a été structuré en fonction de leur type, si ce sont des associations, des musées ou des sites miniers, puis chacun trié relativement au département et à la région auxquels ils font partis ;
- un troisième et dernier scénario qui concerne le thesaurus qui avait était établi au préalable dans le projet ANR-MémoMines.

La prochaine étape de cette mission de stage sera de proposer une cartographie des acteurs et des lieux, grâce à l'utilisation de l'outil de cartographie GEPHI. La cartographie des contenus permet de visualiser les éléments pertinents du corpus et les liens existants entre eux, mais surtout de voir l'architecture de l'information dans ce corpus. Le logiciel Tropes possède déjà ce genre de fonctionnalité mais le but est de voir plus largement les informations dans leur ensemble et pas seulement du point de vue d'une seule référence ou d'une seule relation. La cartographie permet d'établir un plan du contenu du corpus et peut donner l'opportunité de naviguer dans ce contenu et de proposer des itinéraires qui n'étaient pas forcément visible sur un graphe de Tropes ou une grille d'analyse.

# 5 Références bibliographique

## 5.1 Articles

Ouerfelli Tarek, « Archives audiovisuelles et valorisation du patrimoine à l'ère du numérique », Les Cahiers du numérique, 2015/3 (Vol. 11), p. 9-10. URL : https://www-cairn-info.ressources-electroniques.univ-lille.fr/revue-les-cahiers-du-numerique-2015-3-page-9.htm

Leroux Erick, « Introduction », Management & Avenir, 2014/3 (N° 69), p. 147-151. DOI : 10.3917/mav.069.0147. URL : https://www-cairn-info.ressources-electroniques.univ-lille.fr/revue-management-et-avenir-2014-3-page-147.htm

Robert André Désiré, Bouillaguet Annick, « Introduction », dans : André Désiré Robert éd., L'analyse de contenu. Paris cedex 14, Presses Universitaires de France, « Que sais-je ? », 2007, p. 3-8. URL : https://www-cairn-info.ressources-electroniques.univ-lille.fr/l-analyse-de-contenu--9782130563839-page-3.htm

Krieg Alice, « Analyser le discours de presse », Communication, 2000/1 (Vol. 20), p. 75-97. URL : http://journals.openedition.org/communication/6432

Bénédicte Pincemin. « Sept logiciels de textométrie ». 2018. halshs-01843695. URL : https://halshs.archives-ouvertes.fr/halshs-01843695/document

Lebart Ludovic, Piron Marie, « Pratique de l'analyse des données numériques et textuelles avec Dtm-Vic », 2016, Troisième édition. URL : http://www.dtmvic.com/doc/DTM_Manuel_complet_2016.pdf

Robert André Désiré, Bouillaguet Annick, « Chapitre II. Méthodologie générale de l'analyse de contenu et application à un exemple », dans : André Désiré Robert éd., L'analyse de contenu. Paris cedex 14, Presses Universitaires de France, « Que sais-je ? », 2007, p. 24-46. URL : https://www-cairn-info.ressources-electroniques.univ-lille.fr/l-analyse-de-contenu--9782130563839-page-24.htm

Le Coadic Yves-François, « L'objet : l'information », dans : Yves-François Le Coadic éd., La science de l'information. Paris cedex 14, Presses Universitaires de France, « Que sais-je ? », 2004, p. 5-14. URL : https://www-cairn-info.ressources-electroniques.univ-lille.fr/la-science-de-l-information--9782130547495-page-5.htm

Dalbin Sylvie, « Journée d'étude ADBS. La modélisation : pourquoi l'intégrer dans les systèmes d'information documentaire ? », Documentaliste-Sciences de l'Information, 2003/3 (Vol. 40), p. 226-231. DOI : 10.3917/docsi.403.0226. URL : https://www-cairn-info.ressources-electroniques.univ-lille.fr/revue-documentaliste-sciences-de-l-information-2003-3-page-226.htm

## 5.2 Sites web

Ratinaud, P. « Iramuteq », IraMuTeQ, 2020, URL : http://www.iramuteq.org [page consultée le 10 juin 2021]

Contributeurs à Wikipedia, « TXM », Wikipédia, l'encyclopédie libre, 19 avril 2021, URL : https://fr.wikipedia.org/wiki/TXM [Page consultée le 10 juin 2021]

Contributeurs à Wikipedia, « Hyperbase », Wikipédia, l'encyclopédie libre, 13 mai 2021, URL : https://fr.wikipedia.org/wiki/Hyperbase [Page consultée le 10 juin 2021]

Contributeurs à Wikipedia, « Système d'information », Wikipédia, l'encyclopédie libre, 9 janvier 2021, URL : https://fr.wikipedia.org/wiki/Système_d%27information#Système_d'information_et_application_informatique [Page consultée le 14 juin 2021]

# 6 Annexe

- Modèle de Shannon et Weaver : Théorie de l'information

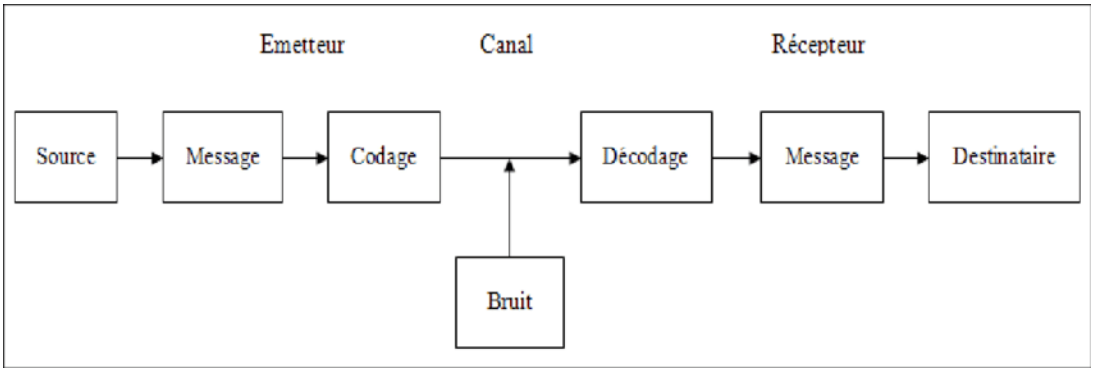



*Figure 20: Modèle de Shannon & Weaver*

- Cours 4 de M2 Structuration de l'information : https://nextcloud.univ-lille.fr/index.php/s/tfr6jSCoZDzctnT

# Résumé et mots-clés

La valorisation du patrimoine représente une certaine forme d'héritage d'une société qui n'existe plus dans les faits mais dont on se souviendra dans les années à venir. Ce mémoire présente le traitement et l'analyse d'un corpus de presse dans le cadre du projet ANR MémoMines qui a été effectué lors d'une mission de stage au laboratoire GERiiCO de l'Université de Lille. Le but de cette mission étant de proposer une grille d'analyse du contenu permettant de repérer les acteurs et les lieux ayant un lien avec le domaine de la mine. La première étape de cette mission a été de nettoyer le corpus d'articles de presse à l'aide du langage XSLT qui est un langage de structuration de l'information. Afin de mener à bien cette mission, une méthodologie a été mise en place, notamment avec l'utilisation du logiciel Tropes, un logiciel d'analyse sémantique. Cette méthodologie a permis de développer plusieurs scénario visant à confirmer des listes préétablies d'acteurs, de lieux ainsi que le thesaurus déjà défini au préalable.

Mots-clés : Patrimoine – Analyse de contenu – Tropes - Modélisation – Structuration de l'information – XSLT