JULIEN VALLIÈRES
Costa Volpino Lombardie Italie +39 377-087-2486
julien.vallieres-gingras@mail.mcgill.ca

v. 2. 20210719

CRIEM-CIRM
Pôle d'analyse de données sociales — Laboratoire d'analyse des discours et des récits collectifs
Projet : Les récits de la faim à Montréal (preuve de concept)
Corpus : textes médiatiques
Auteur : Julien Vallières

RAPPORT RÉCAPITULATIF SUR LA CONSTITUTION DU CORPUS

Ce rapport concerne spécifiquement le moissonnage direct des articles de presse sur l'alimentation publiés sur les sites de vingt-trois médias québécois. ~~Il ne concerne pas le moissonnage des articles de presse sur l'alimentation reproduits dans l'entrepôt de données Eureka de Cision, qui a fait l'objet d'un rapport distinct (Rapport sur l'utilisation du service Eureka), déposé le 22 septembre 2020.~~ Il décrit les étapes de constitution d'un sous-corpus du corpus de textes médiatiques transmis à Pascal Brissette le 25 février 2021, dans le cadre du projet Les récits de la faim à Montréal. En tout, ce travail s'est échelonné sur vingt mois. Deux autres rapports, complétés d'un addenda, déposés les 15 août et 17 août et le 5 septembre 2019, décrivent les étapes initiales du travail exécuté. Pour la description des étapes du travail de relevé bibliographique, préalable au travail de moissonnage décrit ici, on consultera les rapports d'étape reproduits en annexe. À titre de rappel, le mandat original, qui ne concernait que le relevé bibliographique, et dont les termes avaient été fixés lors d'une rencontre le 13 juin 2019, avait été modifié le 30 août de la même année, pour élargir la période du relevé, puis subséquemment, pour la prolonger encore un peu, puis pour inclure les tâches associées au moissonnage et au prétraitement des articles pour leur analyse assistée par ordinateur. L'énoncé du mandat ne contenait pas d'indications méthodologiques, et s'est poursuivi sans précisions ultérieures de cette nature. Un aspect du mandat, par conséquent, dut consister à explorer des méthodes de travail et des solutions logicielles. Le lecteur notera que le présent document ne se conçoit pas indépendamment de son complément, un dossier contenant cinq sous-dossiers, contenant chacun un ensemble de fichiers, dont les copies ont été versées dans un dossier partagé en ligne sur Google Drive (Rapport sur le moissonnage direct - Fichiers).

1. Durée

Entamé à l'automne 2019, après une première phase exploratoire au cours de l'été précédent, le travail de constitution du sous-corpus médiatique des articles de presse sur l'alimentation publiés en ligne sur les sites de vingt-trois médias québécois a été réalisé au cours de quatre périodes de travail distinctes, deux périodes de relevé bibliographique au cours desquelles les textes ont été identifiés et leurs adresses URL recueillies et classées, en novembre 2019 et en juin 2020, une période de moissonnage (*scraping*), soit l'extraction des textes et leur agrégation, en décembre 2020, suivi d'une période de prétraitement des données textuelles sous forme de base de données à fin de leur préparation à l'analyse par traitement informatique, de janvier à février 2021.

~~On trouvera, en annexe, un tableau détaillant sommairement le travail en son entier, compilant les heures, incluant le travail non décrit dans le présent rapport.~~

2. Sources médiatiques

L'élargissement de l'empan de la collecte, décidé en septembre 2019, porté à la période 2005-2020 (plus précisément 2005-01-01 au 2020-05-31), a entraîné un resserrement du nombre de sources

JULIEN VALLIÈRES
Costa Volpino Lombardie Italie +39 377-087-2486
julien.vallieres-gingras@mail.mcgill.ca

médiatiques. La multiplication par cinq du nombre de textes à extraire, le corpus brut atteignant désormais près de 50 000 textes, rendant superflue l'intégration d'ensemble de textes de l'ordre de quelques dizaines. Dix-huit sources médiatiques identifiées à l'été 2019 ont été retenues, seize desquelles ne se trouvent pas sur Eureka. Deux s'y trouvent, *le Journal de Montréal* et *24 heures*, mais dont les archives, pour la période, sont incomplètes sur Eureka (plus de détails ci-après). Une source médiatique non identifiée en 2019 a été ajoutée en 2020. Le réaménagement des archives des textes publiés par *le Journal de Montréal* a entraîné l'ajout, quoique marginal, de quatre autres sources. Deux sources, Gaïa presse puis Vice (Québec), ont cessé leurs activités au cours de la période, soit en 2019. Une autre, HuffPost (Québec), a cessé ses activités après la période, en 2021.

Suivant la méthodologie exposée dans les rapports d'étape, les textes ont été repérés en interrogeant l'index des sites au moyen du moteur de recherche Google, prenant appui sur ses outils de filtrage, dont les résultats ont pu être compilés et extraits à l'aide de l'outil logiciel Google SERPs Extractor (Chris Ainsworth), combiné à l'utilisation de l'extension logicielle gInfinity (Chrome Web Store). Des requêtes par mots-clés ou chaînes de caractères ont été employées afin de repérer les textes. Pour les médias dont le propos est de portée nationale, ces requêtes ont combiné le nom de la ville, *montreal*, à l'un ou l'autre des termes *alimentation, bouffe, food*. Pour les médias dont le propos est de portée locale, soit par restriction, les médias proprement montréalais, la mention du nom de la ville a été abandonnée, et les requêtes ont été faites avec les termes *alimentation, bouffe, food* seulement. Considérant que de façon générale, les textes publiés dans les médias montréalais concernent la ville et sa communauté.

Les vingt-trois sources médiatiques dont les sites ont été moissonnés, leur contribution au corpus et les bornes de publication de cette contribution sont :

média	*période*	*textes*			
24 heures	2015-2020	26	MTL Blog	2013-2020	1713
Baron mag	2009-2020	1296	Narcity (Montréal)	2015-2020	950
Bible urbaine	2013-2020	90	Nightlife	2008-2020	891
Billie	2018	1	Pèse sur start	2017	2
CTV News (Montreal)	2008-2020	2239	Silo 57	2017-2020	214
Daily Hive (Montreal)	2016-2020	731	Ton barbier	2013-2020	200
Gaïa presse	2007-2019	654	Ton petit look	2012-2020	441
HuffPost (Québec)	2012-2020	1870	TVA Nouvelles	2005-2020	1028
Journal de Montréal	2012-2020	1134	Un point cinq	2017-2020	367
Sac de chips	2015-2020	36	Vice (Canada)	2005-2019	1506
Montreal Gazette	2005-2020	2331	Voir	2005-2020	879
Montréal.TV	2012-2020	253			
				total	18 852

3. Procédures d'extraction

Des relevés bibliographiques ont été produits pour chacun des médias, combinant dans des tableurs Excel (.xlsx), pour chaque texte repéré, nom de la source, année de publication, titre du texte, texte d'accroche (*anchor text*) et adresse URL de publication. En un second temps, depuis ces tableurs, les adresses URL ont été exportées dans des fichiers tabulaires élémentaires (.tsv) (dossier - URL), puis manipulées dans le logiciel d'édition BBEdit (Bare Bones Software), et insérées, par blocs, dans des patrons de requête (*sitemaps*). Produits à l'aide de l'extension logicielle Web Scraper (Web Scraper) du navigateur Chrome, ces patrons de requêtes une fois *remplis* ont ensuite été

reversés dans Web Scraper au moyen de la fonction d'importation du logiciel. La procédure de moissonnage comme telle a été ensuite prise en charge par le logiciel. Pour chaque média, des tableurs ont été produits affichant les résultats, dont les données ont été copiées, agrégées lorsque les textes étaient trop nombreux, puis versées dans un espace de travail distinct dans le logiciel de manipulation de données OpenRefine (OpenRefine).

Ces patrons de requêtes (dossier - sitemap) sont exportables au format JSON. Pour maximiser les résultats du moissonnage, ils combinent de quinze à vingt requêtes d'extraction, fondées sur les balises HTML des pages. Un échantillon de pages présentant des variations a été étudié pour mettre au point chacun des patrons de requêtes. Cet échantillonnage a été rendu particulièrement nécessaire en raison de l'ampleur de la période de cueillette (2005-2020), au cours de laquelle les sites des médias ont fait l'objet de mises à jour, quand ce ne sont pas des migrations complètes, leurs contenus archivés réunis ou réorganisés, sources d'irrégularités qu'il s'est agi de ne pas échapper, autant que faire se peut.

~~3.1. Le cas du *Journal de Montréal*~~

~~Il a été signalé, dans le Rapport sur l'utilisation du service Eureka, qu'un problème se présentait pour le moissonnage de nombreux textes du *Journal de Montréal* et du *24 heures* archivés dans le dépôt numérique de Cision. Une première tentative de résoudre le problème n'avait pas donné de résultat. Tandis qu'était entrepris ce volet-ci du moissonnage, le problème a été réexaminé. Le problème était le suivant : les archives d'un millier et demi d'articles ne contenaient que le liminaire du texte, suivi d'une URL dirigeant vers un site extérieur, en l'occurrence celui du *Journal de Montréal*. Pour compliquer les choses, par suite d'une réorganisation des contenus journalistiques en ligne de Quebecor, le quart des URL dirigeaient vers des pages dont le contenu avait été déplacé sur de tiers sites spécialisés, adressés à des auditoires ciblés, soit les sites *Silo 57*, *Sac de chips*, *Pèse sur start*, *Billie*.~~

~~Le problème a été résolu en deux temps. Premièrement, les textes dont Eureka ne conserve que les titre et liminaire ont été moissonnés de nouveau, pour en extraire les URL de redirection. Ces URL ont ensuite été compilées à leur tour dans des fichiers tabulaires élémentaires, comme celles obtenues en interrogeant l'index des sites. Deuxièmement, à l'aide de Web Scraper, un super patron de quarante requêtes a été produit pour moissonner les textes sous tous les cas de figure, selon le site de destination de la simple ou double redirection. Ont pu ainsi être moissonnés les textes que des requêtes par mots-clés avaient permis le repérage dans Eureka, et ce malgré que ces textes n'aient pas été archivés sur Eureka, en moissonnant directement les sites du *Journal de Montréal* et de ses dépendants.~~

4. Procédures de prétraitement

Le prétraitement des données textuelles moissonnées a été réalisé à l'aide du logiciel OpenRefine. Pour chacune des sources médiatiques, les données extraites ont été versées dans un tableur, structurées et nettoyées. Les résultats des requêtes concurrentes ont été comparées, les exceptions, identifiées, les valeurs, selon le besoin, divisées ou concaténées. Dans le cas des données de deux sources, soit *le Journal de Montréal* et *TVA Nouvelles*, le traitement en bloc de l'ensemble des données obtenues provoquant un ralentissement logiciel notable quoiqu'inexpliqué, ces ensembles

de données ont été fractionnés, et les opérations de prétraitement, répétées pour chaque sous-ensemble.

Publiées en plein texte courant, les données textuelles moissonnées n'ont nécessité que de parcimonieuses corrections. Trois principaux cas de figure se sont présentés. Soit qu'une utilisation sous optimale des balises HTML par les éditeurs ait été compensée par une surutilisation des retours de chariots et de signes de ponctuation divers. Soit que la procédure de moissonnage ait généré des caractères typographiques additionnels superflus pour marquer la mise en forme originelle des textes. Soit que les textes, sur le site même où ils ont été moissonnés, présentaient des caractères inconnus, résultat probable d'une migration de contenu antérieur par l'éditeur, ou d'un encodage vestigial.

La structuration des données textuelles moissonnées a requis un nombre élevé de manipulations, dû en partie à l'inexpérience que j'avais dans l'utilisation du logiciel OpenRefine. L'utilisation des balises HTML montrant des irrégularités nombreuses, en particulier sur les sites des éditeurs qui ne recourent pas à une solution logicielle commerciale, et à plus forte raison sur une longue période, durant laquelle l'éditeur a procédé à des réorganisations répétées de ses contenus, pour reconstituer le corps du texte, il a fallu se rabattre sur des requêtes de moissonnage gourmandes, qui ont concaténé sous des balises mères les textes, ne tenant pas compte des balises filles. Dans d'autres cas, spécialement pour les textes plus anciens — comme ceux du *Canal Argent*, archivés sur le site de *TVA Nouvelles* —, mais pas seulement — c'était le cas aussi des textes de *Gaïa presse* —, s'est présenté le problème inverse, invoqué dans le paragraphe précédent, de l'utilisation sous optimale des balises, les noms des auteurs, par exemple, ou des énoncés biographiques les concernant, les titres, sous-titres ou intertitres, voire les indications de rubricage, le chapeau, les légendes des illustrations, se trouvant inclus dans la balise de corps de texte. L'opération la plus délicate et la plus chronophage aura consisté par conséquent à repérer et capturer des chaînes de caractères, du mot à l'énoncé complexe, pour les extraire de valeurs semi-structurées, et remplir les champs appropriés.

Le format des valeurs contenant les noms d'auteur a été standardisé. Le regroupement (*clustering*) des entités énonciatrices a été réalisé — par entités énonciatrices, on entend les auteurs, mais aussi les noms des sources lorsque les textes provenaient d'une agence de presse, d'un autre média ou d'un organisme et le mentionnaient.

Les opérations sur les données ont été effectuées au moyen d'expressions régulières (regex) en langage GREL, le langage de manipulation de données du logiciel OpenRefine. Le logiciel, qui permet, à tout moment, de remonter la chaîne des opérations, permet ainsi de la reconstituer. Pour chacun des ensembles et sous-ensembles de données prétraités, a été produit un tableur contenant l'historique des opérations GREL (dossier - opérations GREL). Ces chaînes d'opérations reconstituant l'historique de manipulation des données, traduits dans le langage d'interopérabilité JSON, ont été exportées et versées, sous cette autre forme, au dossier (dossier - opérations JSON).

Volontiers fastidieux, le nettoyage de données semi-structurées implique la capture, le déplacement, la substitution de nombreuses chaînes de caractères, souvent récurrentes. Pour mener à bien ce travail, il n'y a d'autre choix que de compulser les données. Durant ce travail, des relevés ont été constitués, en prévision de la formulation d'opérations de manipulation des données. Ces relevés étaient conservés durant le travail dans des fichiers textes élémentaires intitulés Carnets de

correction. À titre d'exemple, les contenus de six de ces carnets ont été versés au dossier (dossier - carnets). On aura soin de les consulter pour ce qu'ils sont, des outils de travail, transitoires, qui n'ont pas été produits dans l'intention de conserver la trace de toutes les manipulations. En l'état où ils sont, néanmoins, quoiqu'incomplets, ils offrent un coup d'œil complémentaire dans l'atelier virtuel du *data scientist*.

En l'absence de consignes balisant le prétraitement, pour ne pas faire obstacle à des analyses plus fines éventuelles, et dans la perspective ~~désormais improbable~~ d'une diffusion publique du jeu de données produit, le nettoyage visa un état de données très propre. Il y a lieu de réévaluer, en aval, ce choix méthodologique.

4.1. Format RIS

Les données textuelles du sous-corpus ont été structurées en reprenant les balises à doubles lettres du format RIS, un format standard d'interopérabilité bibliothéconomique largement utilisé. Ce choix a été suggéré par le recours par Cision dans le dépôt numérique Eureka au format RIS pour l'organisation des métadonnées que fournit le service sur les archives médiatiques que le dépôt contient, et motivé par un souci de standardisation, rendu nécessaire par la réunion ultérieure des sous-corpus obtenus par moissonnage.

Un fichier texte élémentaire (.rtf) a été produit contenant et la description des champs de données du corpus sous les balises RIS correspondantes de même que quelques éléments de description du corpus (CRIEM_recitsfaim_googlecat_meta).

JULIEN VALLIÈRES
Costa Volpino Lombardie Italie +39 377-087-2486
julien.vallieres-gingras@mail.mcgill.ca

RAPPORT RÉCAPITULATIF
SUR L'UTILISATION DU SERVICE EUREKA (CISION)

Ce rapport concerne l'utilisation du service de veille médiatique Eureka de Cision, une base de données contenant des archives textuelles de milliers de médias canadiens. Il décrit les étapes de constitution d'un corpus de textes médiatiques, partagé avec les autres membres de l'équipe le 16 septembre 2020, dans le cadre du projet Les récits de la faim à Montréal. En tout, ce travail s'est échelonné sur un an. Puisque l'utilisation du service Eureka n'avait pas été évoquée dans les deux rapports précédents et leur addenda, précisons qu'elle a été rendue nécessaire pour répondre à la demande d'accroître la période couverte par la recherche, pour couvrir, au lieu que les années 2017-2018, les années 2000-2019, demande transmise par courriel par Pascal Brissette le 30 août 2019. La période 2000-2004 étant peu couverte par le service Eureka, qui élargit la veille médiatique à un nombre sensible de sources en 2004-2005, il a été convenu, ultérieurement, de restreindre la période aux années 2005-2019. Pour étudier l'impact de la pandémie de coronavirus, la cueillette a inclus également les premiers mois de l'année 2020 (jusqu'en mai, inclusivement).

La constitution du sous-corpus médiatique Eureka a été réalisée au cours de trois périodes de travail distinctes, dont deux périodes d'identification des textes et cueillette bibliographique, en octobre 2019 et juin 2020, et une période de moissonnage (*scraping*) et prétraitement des textes en août et septembre 2020.

Facilitée par le moteur de recherche inclus dans son service, l'identification des textes s'est faite à l'aide de requêtes de mots (chaînes de caractères), après sélection de douze sources médiatiques ou sous-ensembles de sources, pour la période 2005-01 à 2020-05, ou des périodes plus courtes, selon les bornes de la période couverte par le service pour une source particulière. Ces sources sont :

	période		période
La Presse, Cyberpresse, La Presse+	2005-2020	Journaux de quartier montréalais fran.	2005-2020
Journal de Montréal	2006-2020	Journaux de quartier montréalais engl.	2005-2015
24 heures Montréal	2012-2020	Radio-Canada	2005-2020
Le Devoir	2005-2020	Radio-Canada Montréal	2005-2020
Journal Métro Montréal	2005-2020	CBC	2005-2020
L'Actualité	2005-2020	CBC Montreal	2005-2020

Comme pour l'identification des textes à partir de l'indexation des sites par Google, les requêtes utilisées ont employé, selon le cas, l'un ou l'autre des termes *Montréal, alimentation, bouffe, food,* ou leur combinaison.

Les résultats des requêtes ont ensuite été exportés au format RIS, à l'aide d'une fonction offerte par le service. Le format RIS est un format bibliographique standard. Les descriptions bibliographiques fournies par le service Eureka contiennent seize champs, dont un pour l'adresse URL de l'archive du texte médiatique. Le format d'extraction est un fichier compatible TXT. Pour obtenir un fichier tabulaire, une procédure ultérieure simple de transposition des lignes en colonnes est nécessaire. Lorsque les résultats étaient trop nombreux, la période a été segmentée et les requêtes répétées. Les résultats ont ensuite été concaténés à l'aide d'une commande dans Terminal (Mac). Des copies des douze fichiers distincts au format RIS obtenus ont été conservées.

La procédure de moissonnage a été réalisée à l'aide de l'extension logicielle Web Scraper du navigateur Google Chrome. En dépit qu'elle portât sur les archives de nombreuses sources, la procédure a été facilitée par l'uniformisation de l'affichage des archives du service Eureka. Les

fichiers au format RIS ont été convertis au format tabulaire, en effectuant la transposition des seize lignes en autant de colonnes, une pour chaque champ de la description bibliographique. La colonne contenant les adresses URL des archives des textes médiatiques a ensuite été isolée et les adresses URL ont été extraites puis elles ont été versées dans des fichiers TXT, une par ligne. Sur la liste obtenue ont ensuite été appliquées quelques traitements simples, pour insérer les adresses dans une ligne de code servant de parcours de moissonnage au logiciel Web Scraper, et cette ligne de code importée dans le logiciel. Lorsque le nombre d'adresses était jugé trop élevé, après une ou deux expériences malheureuses, les listes ont été fractionnées, de manière à répéter plusieurs fois la procédure, un millier de textes à la fois. La procédure de moissonnage a porté sur quatre champs, pour autant de rubriques : le corps du texte, le chapeau, soit le segment textuel mis en évidence en tête du texte, l'encadré, soit un texte d'accompagnement du texte principal, la légende de l'illustration. Les textes ont été moissonnés en récupérant l'information relative à leur segmentation (saut de paragraphe).

Les résultats ont ensuite été copiés, versés dans le logiciel BBEdit, et le cas échéant, concaténés, de sorte à n'obtenir qu'un fichier par source. Les erreurs de moissonnage ont ensuite été isolées et leurs adresses URL extraites, et le moissonnage, repris, pour ces seules adresses, sur un rythme de requête plus lent, afin de récupérer les textes qui avaient échappé à la procédure la première fois. Ces résultats ont été combinés aux résultats précédents. Les textes pour lesquels la double procédure de moissonnage a échoué, pour les sources *Journal de Montréal* et *24 heures*, ont été identifiés, leurs adresses URL extraites et réunies, puis les résultats incomplets (null) retirés. Une tentative de récupérer autrement une partie de ces textes, sans passer par le service Eureka, au moyen d'adresses URL moissonnées dans les résultats incomplets, n'a pas donné de résultat. Au total, le moissonnage des archives médiatiques du service Eureka a permis de composer un sous-corpus préliminaire de 27 667 textes médiatiques, articles et reportages se rapportant, de près ou de loin, à l'alimentation à Montréal. Les résultats du moissonnage ont été à leur tour concaténés de sorte à n'obtenir qu'un fichier. Soit les résultats suivants par sources :

	nb de txt		nb de txt
La Presse, Cyberpresse, La Presse+	6383	Journaux de quartier montréalais fran.	1475
Journal de Montréal	3361	Journaux de quartier montréalais engl.	3975
24 heures Montréal	1239	Radio-Canada	1390
Le Devoir	2706	Radio-Canada Montréal	453
Journal Métro Montréal	3261	CBC	1025
L'Actualité	289	CBC Montreal	2110

Le prétraitement des résultats, en vue de leur ingestion éventuelle dans la base de données, a été réalisé à l'aide du logiciel OpenRefine. Le texte obtenu en moissonnant les archives du service Eureka est très propre et sa préparation n'a exigé aucune correction. Seules les marques additionnelles générées pour conserver l'information relative à la segmentation des textes ont dû être nettoyées, ainsi que de rares indications vestigiales de saut de ligne. Pour garder trace de la segmentation du texte, de façon uniforme, le caractère de la lettre allemande eszett (ß) est utilisé pour signaler un saut de ligne.

Le moissonnage avait porté sur quatre champs, excluant les données déjà contenues dans les descriptions bibliographiques au format RIS fournies par Eureka. Afin de conserver ces données bibliographiques, il a fallu convertir au format tabulaire les fichiers au format RIS à l'aide d'OpenRefine, procéder à la transposition des lignes en colonnes, puis générer un identifiant

unique, dérivé de son adresse URL, pour chaque archive. Pour chacune des archives moissonnées également, un identifiant unique a été généré, dérivé pareillement de son adresse URL, soit la même adresse que la précédente. Il s'est agi ensuite de croiser les contenus des tables, à l'aide d'OpenRefine, de façon à ce que pour chaque archive médiatique moissonnée, une description bibliographique en seize rubriques a été ajoutée, indiquant son titre, son auteur, sa source, *etc.*

Un fichier texte au format RTF a été produit pour décrire de manière succincte le corpus.

JULIEN VALLIÈRES
Costa Volpino Lombardie Italie +39 377-087-2486
julien.vallieres-gingras@mail.mcgill.ca

Annexe 1. Détail des heures travaillées à la constitution du corpus du pilote des Récits de la faim en date du 25 février 2021

	20190629 *au* 20190928	20191007 *au* 20191231	20200127 *au* 20200430	20200501 *au* 20201230	20210101 *au* 20200225	
Champ sémantique	4	-	-	-	-	4
Cueillette bibliographique	17	44	-	-	-	61
Moissonnage URL	40	46	-	13	-	99
Moissonnage contenus	-	-	-	109	3	112
Prétraitement	-	-	-	7	168	175
Méthodologie	6	34	12	23	15	90
					total	541

JULIEN VALLIERES
Montréal Québec Canada 514-273-8613
julien.vallieres-gingras@mail.mcgill.ca

20190815

ANNEXE 1

CRIEM-CIRM
Observatoire des récits de Montréal
Projet : Les récits de la faim à Montréal
Corpus : textes médiatiques
Auteur : Julien Vallières

RAPPORT D'ÉTAPE
50 HEURES

Lors d'une rencontre, le 13 juin 2019, entre Pascal Brissette et moi, il a été entendu de consacrer l'intégralité d'un contrat de 104 heures commençant le 29 juin et se terminant le 28 septembre 2019 à la collecte d'adresses URL pour, ultérieurement, moissonner (*scrap*) les textes qui s'y trouvent publiés, en vue de constituer un sous-corpus du corpus du projet « Les récits de la faim à Montréal ». — Il s'agit de rassembler un ensemble de textes médiatiques parlant de l'alimentation à Montréal. Le mot-clé est ici *alimentation*. Il s'agit donc de reconstituer, pour 2017-2018, le discours sur l'alimentation à Montréal.

La procédure suivie pour mener la collecte bibliographique dans le cadre du projet sur les discours entourant le changement de nom des Redmen, qu'il était convenu de reprendre pour ce projet-ci, s'est avérée trop laborieuse pour des résultats de cueillette préliminaires trop importants. Cette procédure visait l'exhaustivité et consistait, après lecture en diagonale des textes identifiés à l'aide d'un moteur de recherche (Google), à retranscrire manuellement les adresses URL des pages où ils sont publiés, leurs titres, ceux des médias qui les hébergent, les noms de leurs auteurs, les titres des sections dans lesquelles ils étaient publiés, ainsi qu'une citation parfois, selon cinq sous-ensembles.

Cette procédure écartée, après un peu de taponnage, une autre lui a été préférée, plus adaptée aux objectifs du travail. Cette procédure-ci consiste en premier lieu à moissonner les résultats de recherche eux-mêmes, à les nettoyer et à dresser des tables annuelles par sources contenant deux champs principaux, la liste des adresses URL glanées à l'aide du moteur de recherche, et la liste des titres des textes publiés à ces adresses. En second lieu, à la lecture des titres, un tri préliminaire sera effectué, pour écarter les textes qui, à l'examen, ont peu à voir avec le sujet de l'alimentation (les « faux positifs »).

À noter que le changement de procédure permet de rétablir l'objectif initial de collecter les textes médiatiques publiés en 2017 et 2018, plutôt que 2018, uniquement, restriction qui m'avait semblée nécessaire d'abord. Par conséquent, les résultats obtenus jusqu'à maintenant couvrent les deux années. On pourra, si on le préfère, dans les étapes ultérieures, restreindre à nouveau la période.

Les sites d'information quotidiens suivants ont été dépouillés :

journaldemontreal.com
ledevoir.com
lapresse.ca
ici.radio-canada.ca
tvanouvelles.ca
journalmetro.com
quebec.huffingtonpost.ca
lactualite.com
montrealgazette.com (anglais)
ctvnews.ca (anglais)
cbc.ca (anglais)

JULIEN VALLIERES
Montréal Québec Canada 514-273-8613
julien.vallieres-gingras@mail.mcgill.ca

Un aperçu du travail effectué, le produit du moissonnage des résultats du dépouillement, par année, par source, nettoyé mais non encore trié, est versé dans un dossier partagé sur Google Drive (un dossier Dropbox suivra plus tard), sous trois formats tabulaires de sauvegarde, .xlsx, .csv et .txt :

CRIEM récits de la faim 2017-2018

Pour la suite, il est proposé de dépouiller un choix de sites parmi trois ensembles de sources. Le premier sous-ensemble regroupe les sites d'information alternatifs suivants, basés à Montréal :

voir.ca
vice.com/fr_ca
nightlife.ca
narcity.com
montreal.tv
tonpetitlook.com
tplmoms.com
tonbarbier.com
mauditsfrancais.ca
novae.ca
gaiapresse.ca
mtlblog.com (anglais)
dailyhive.com/montreal (anglais)

Le deuxième sous-ensemble regroupe des sites d'information spécialisés, incluant les sites de magazines culinaires québécois :

actualitealimentaire.com
foodlavie.com
tastet.ca
onsenfood.com
potluckmtl.com
recettes.qc.ca
coupdepouce.com
ricardocuisine.com
cariboumag.com
dinettemagazine.com
lemust.ca
foodbloggersofcanada.com (anglais)

Le troisième sous-ensemble regroupe une sélection de blogues culinaires montréalais, actifs en 2017 et 2018 :

fraichementpresse.ca
montreal-addicts.com
familleettofu.com
christelleisflabbergasting.com
emiliemurmure.com
barbaragateau.com
uneparisienneamontreal.com
curiositesetgourmandises.com
unemerepoule.com
lecoupdegrace.ca
mestrouvailles.ca
boucheesdoubles.net
cerisesetgourmandises.com
nutritionnisteurbain.ca
marielenfer.com
mespetitesrevolutions.com
urbainecity.com
turquoise-blog.com (bilingue)

La question des recettes. Doit-on tenter d'écarter les recettes ? Si une majorité de sources en publie, les blogues culinaires, en plusieurs cas, publient presque exclusivement des recettes.

La question des critiques de restaurant. En grand nombre, veut-on tenter d'écarter préalablement aussi les critiques de restaurant ?

La question de l'anglais. Doit-on moissonner les sources anglophones ? Du côté de l'analyse de texte, le bilinguisme pose des problèmes spécifiques. Néanmoins, pour la constitution des listes bibliographiques, puisque cette question n'en pose pas beaucoup, à cette étape, j'ai inclus quelques sites anglophones.

Les mots-clefs. Si on exclut la sélection préalable des sites et le choix des années, pour les sources francophones, la coprésence de deux mots-clefs a représenté la seule contrainte de recherche, soit *Montréal* et *alimentation*. Pour les sources anglophones, *alimentation* ne produisant presque pas de résultats (« Alimentation Couche-Tard » principalement) et n'ayant pas d'équivalent strict, un autre terme, *food*, lui a été substitué, accompagné de *Montréal* une fois de plus. Des requêtes exploratoires ont cependant produit des résultats curieux : le terme *food* est largement employé dans les textes français. Qui plus est, les résultats obtenus sont en maints cas complémentaires à ceux obtenus en employant le terme *alimentation*. Sur certaines sources, par exemple *nightlife.ca*, une source francophone malgré le titre emprunté à l'anglais, une requête avec le terme alimentation produit peu de résultats, tandis qu'une requête avec le terme *food* en produit beaucoup. Que faire en ce cas ?

Pour finir, bien que ce ne soit pas un texte médiatique, je partage ce lien vers la description d'une exposition tenue à l'Écomusée du fier monde du 18 mai 2017 au 4 février 2018 :

Nourrir le quartier, nourrir la ville

JULIEN VALLIERES
Montréal Québec Canada 514-273-8613
julien.vallieres-gingras@mail.mcgill.ca

ANNEXE 2

CRIEM-CIRM
Observatoire des récits de Montréal
Projet : Les récits de la faim à Montréal
Corpus : textes médiatiques
Auteur : Julien Vallières

RAPPORT D'ÉTAPE
50 HEURES
ADDENDA

Élaboration de la question des mots-clefs, exemple à l'appui.

Comme mentionné, des requêtes exploratoires ont produit des résultats qui suggèrent une réflexion sur les mots-clefs utilisés afin de repérer les textes candidats au moissonnage. En interrogeant l'index de sites francophones, des requêtes distinctes combinant les termes *montreal–bouffe* et *montreal–food* ont produit des résultats complémentaires aux résultats obtenus au moyen de la combinaison *montreal–alimentation*. Ces requêtes ont produit des listes ou bien comparables en longueur, ou bien plus longues que les listes obtenues au moyen de cette dernière combinaison.

À titre d'exemple, en interrogeant à l'aide de Google l'index de l'édition québécoise du site *vice.com*, au moyen de trois requêtes employant les combinaisons mentionnées ci-devant, on obtient les résultats suivants, consolidés et classés par année de référence :

Vice (édition québécoise) 2017

Vice (édition québécoise) 2018

Pour l'association des résultats avec les termes des requêtes au moyen desquelles ils ont été obtenus, voir tout à droite les colonnes E, F, G, H.

JULIEN VALLIERES
Montréal Québec Canada 514-273-8613
julien.vallieres-gingras@mail.mcgill.ca

ANNEXE 3

CRIEM-CIRM
Observatoire des récits de Montréal
Projet : Les récits de la faim à Montréal
Corpus : textes médiatiques
Auteur : Julien Vallières

RAPPORT D'ÉTAPE
90 HEURES

Exposé de l'état du travail à l'approche de la quatre-vingts-dixième heure. On tâchera de ne pas répéter ce qui a précédemment été dit.

Le moissonnage des adresses URL des localisations, en ligne, des textes médiatiques se rapportant à l'alimentation publiés en 2017 et 2018 à Montréal s'est poursuivi. En tout, les sites internet de trente sources médiatiques ont été dépouillés, qui peuvent être partagés en trois groupes, soit des sites d'information quotidiens, des sites d'information alternatifs et des magazines culinaires et sites spécialisés. Du nombre, vingt-quatre sources publient en langue française et six, en langue anglaise. Sélectionnées pour moitié en raison de leur situation dans le paysage médiatique, moitié pour l'abondance des résultats obtenus à l'occasion de recherches exploratoires, ces sources sont, ci-dessous accompagnés du nombre de textes identifiés pour chacune des années de référence :

	2017	2018		2017	2018
journaldemontreal.com	757	756	lemust.ca	58	51
journalmetro.com	386	421	tonpetitlook.com	38	29
ledevoir.com	387	382	montreal.tv	43	22
ici.radio-canada.ca	332	377	fr.chatelaine.com	24	40
tvanouvelles.ca	304	359	vice.com/fr_ca	28	23
voir.ca	316	264	cariboumag.com	19	26
lactualite.com	249	311	coupdepouce.com	20	17
baronmag.com	263	279	recettes.qc.ca	5	10
quebec.huffingtonpost.ca	259	274	dinettemagazine.com	nd	12
lapresse.ca	202	296	mtlblog.com *ang.*	296	299
gaiapresse.ca	286	193	cbc.ca *ang.*	298	296
narcity.com	148	142	montrealgazette.com *ang.*	287	303
tastet.ca	182	92	ctvnews.ca *ang.*	284	267
nightlife.ca	130	132	dailyhive.com/montreal *ang.*	295	99
tonbarbier.com	67	43	foodbloggersofcanada.com *ang.*	13	21

Au total, pour les années 2017 et 2018, on obtient 4503 et 4551 textes de langue française ainsi que 1473 et 1285 textes de langue anglaise. Soit 5976 et 5836 textes dans les deux idiomes.

À noter que les résultats préliminaires produits en interrogeant le site du *Journal de Montréal* affichent un taux de pertinence plus faible, en raison à la fois du nom du quotidien, qui rend inopérante la discrimination des résultats à l'aide du nom de la ville, et de l'organisation des pages du site, où le corps du texte d'un article donné est publié sur le même plan que les titres de nombreux articles, affichés sur les côtés.

Ces résultats ont été obtenus à l'aide de requêtes employant, selon le cas, l'un ou l'autre des termes *Montréal, alimentation, bouffe, food*, ou leur combinaison. Les médias généraux d'information, quotidiens ou alternatifs, ont été interrogés à l'aide de trois requêtes combinées, sous la forme *Montréal–alimentation*. Lorsque c'est le cas, on trouvera, dans les trois colonnes situées le plus à droite des tables de résultats, les mots-clefs qui m'ont permis de les identifier précédés d'un croisillon (#). Lorsque plus d'une requête a identifié un même texte, les doublons ont été éliminés, en rapportant les mots-clefs multiples sur une même ligne. Pour certains médias étroitement associés à la métropole, l'inclusion de son nom dans les requêtes a été jugée superflue. À l'inverse, le dépouillement des médias spécialisés s'est fait à l'aide de requêtes employant seulement *Montréal*.

Comme mentionné dans le rapport de cinquante heures, le produit du moissonnage des résultats du dépouillement, par année, par source, nettoyé mais non encore trié, est versé à un dossier dont des copies sont partagées à la fois sur Google Drive et Dropbox :

CRIEM récits de la faim sur Google Drive

CRIEM récits de la faim sur Dropbox

Soixante fichiers tabulaires produits, un par source, trente par année, contiennent l'ensemble des adresses moissonnées. Un tri que je proposais de réaliser pour écarter les textes qui, à l'examen des titres glanés, ont manifestement peu à voir avec le sujet de l'alimentation n'a pas encore été effectué. Vu la composition de l'équipe, ce tri pourrait être réalisé ultérieurement, au moyen d'algorithmes.

En attendant que soit décidée la question qui précède, suivant les dernières directives transmises, je travaillerai à l'établissement d'une bibliographie des fictions narratives (roman, nouvelle, conte, récit) publiées entre 2017 et 2019 inclusivement et qui mettent en jeu l'espace montréalais.

Avant quoi, je rédigerai une note concernant l'évaluation des implications de l'élargissement de l'empan chronologique du corpus de manière à couvrir la période allant de 2000 à 2019.