Conception d'accès thématiques et d'initiatives de Knowledge Management depuis un corpus structuré : text-mining, segments répétés et communautés de pratiques au service de l'accès à l'information et du partage de connaissances

Illan Marc Obadia

CONSERVATOIRE NATIONAL DES ARTS ET MÉTIERS

Equipe pédagogique Stratégies

INTD

MÉMOIRE pour obtenir le Titre enregistré au RNCP

"Chef de projet en ingénierie documentaire et gestion des connaissances"

Niveau I

Présenté et soutenu par

Illan Marc Obadia

Le 19 Décembre 2018

Conception d'accès thématiques et d'initiatives de Knowledge Management depuis un corpus structuré

text-mining, segments répétés et communautés de pratiques au service de l'accès à l'information et du partage des connaissances

Jury :

M. Loïc Lebigre, PAST, Directeur de Mémoire,
M. Frédéric Erlos, Responsable de la coordination, textes de gouvernance, délégation et réclamation chez Crédit Agricole S.A. Docteur en sciences du langage

Promotion 48 (2017-2018)

Remerciements

Reprenant des études au courant de ma carrière, l'Institut National des sciences & Techniques de la Documentation m'a offert l'opportunité de renforcer et découvrir des savoirs, concernant les métiers de l'ingénierie documentaire et de la gestion des connaissances.

Que je puisse exprimer ma reconnaissance envers l'équipe pédagogique de l'INTD et en particulier à mon Directeur de de Mémoire le Professeur Loïc Lebigre. Je remercie tout autant le Professeur Associé Gonzague Chastenet De Géry, pour m'avoir indiqué et encouragé dans cette démarche.

Je remercie mes collègues de promotion. Nos différences et expériences ont été le creuset d'un melting-pot de savoir-faire passionnant.

Dans le cadre de ce stage, je remercie mes collègues du Secrétariat Général de Crédit Agricole S.A. pour leur bienveillance.

Ma gratitude va également à Frédéric Erlos pour son accueil, la richesse de nos échanges et ses précieux conseils. Je remercie Anne, Alexis, Caroline, Michèle et David.

Je remercie le Professeur Jean-Hubert Blanchet et M. David Pilczer pour leurs précieuses informations sur le secteur financier.

Merci à mon employeur, l'entreprise Modis France et en particulier à ma Directrice, Mme Stéphanie Porra pour m'avoir accordé le temps nécessaire à la réussite de ce projet.

Enfin, mes pensées vont à ma famille que je remercie pour son soutien et sa patience. Merci à Rose, à Ruth et Jean-Claude, à ma mère et à Sivane. Merci d'avoir cru en ce projet dont la réussite me tenait à cœur.

Notice

OBADIA Illan Marc. Conception d'accès thématiques et d'initiatives de Knowledge Management depuis un corpus structuré. Text-mining, segments répétés et communautés de pratiques au service de l'accès à l'information et du partage de connaissances.

Basés sur des intranets documentaires, les corpus de documents structurés héritent de classification à facettes essentiellement utilisés pour organiser les documents selon leurs types.

Cumulant parfois de nombreuses années de documentations, leur contenu évolue au fil des évolutions contextuelles. La complexité et la contingence des sujets contenus dans les documents peut parfois dérouter les travailleurs de la connaissance.

Via l'analyse de l'audience d'un intranet comprenant un corpus structuré et le text-mining sur son contenu, il est possible de concevoir de nouveaux systèmes d'organisation des connaissances érigés en accès thématiques. L'approche processus réalisée régulièrement permet alors d'adapter l'offre d'accès au contexte de l'organisation consultant le corpus.

Des initiatives complémentaires de gestion des connaissances contribuent à étendre l'info-documentation à l'info-connaissance dans l'objectif d'atteindre l'excellence opérationnelle.

Ce mémoire illustre un cas d'usage professionnel de conception de nouveaux accès thématiques par une utilisation professionnelle de la textométrie, en particulier des segments répétés réalisé au sein du Groupe Crédit Agricole.

En conclusion, il propose un modèle d'indexation automatisé des documents dans des thématiques via l'usage des segments répétés.

Abstract

OBADIA Illan Marc. Based on documentary intranet, the bodies of structured texts benefits from faceted classification essentially used to organize documents according their types.

Sometimes cumulating many years of documentations, their content change through contextual evolution. The complexity and contingency of subjects contained in the documents can sometimes disconcert the knowledge workers.

Through web analytics of a documentary intranet and text-mining on the contents, it is possible to design new knowledge organization system based on thematic accesses. The process-approach carried out regularly, makes possible to adapt the access to the context of the organization consulting the corpus.

Complementary knowledge management initiatives can help to extend the Info-Documentation to Info-Knowledge with the goal of achieving operational excellence.

This thesis illustrates a professional case of designing of new thematic accesses by a professional use of text-mining,in particular, for repeated segments realized within Crédit Agricole Group.

It proposes in conclusion, a model of automatic indexing of documents in themes.

Sommaire

Liste des tableaux

Liste des illustrations

Introduction

L'histoire des grandes entreprises historiques résulte de leur croissance organique et de multiples fusions et acquisitions, ayant eu pour conséquence de les rendre vitales à l'économie auxquelles elles contribuent.

Parallèlement à ces regroupements, le législateur a fortement règlementé leurs activités afin de préserver leur écosystème.

Le système bancaire, pilier de l'économie mondiale, contribue à son essor. Les groupes bancaires, d'une taille critique, sont qualifiés d'institutions financières d'importances systémiques. i.e. l'importance de leur taille et de leurs interconnexions causerait de graves troubles à l'économie en cas de défaillance majeure.

Ces nouveaux défis contribuent à la recomposition régulière de leur organisation, tandis que le législateur et les autorités règlementent et contrôlent leurs activités.

Les banques doivent constamment s'adapter aux contextes économiques fortement concurrentiels entre autres impactés par la transformation numérique.

Assujettis à une inflation règlementaire, les groupes bancaires doivent suivre la législation et être en mesure de répondre à diverses instances de contrôles internationales ayant des pouvoirs de sanctions.

Les enjeux sont économiques, règlementaires, financiers, politiques et digitaux. L'ensemble des strates hiérarchiques de ces entreprises doit comprendre, adapter et partager l'application de diverses directives internationales dans leurs activités.

En réponse à ces défis, l'ingénierie documentaire et la gestion des connaissances permet de mieux valoriser et retrouver l'information règlementaire et ses déclinaisons opérationnelles.

Par de multiples initiatives d'analyse, de classement, de valorisation et de représentation de la documentation et de l'information, il devient possible d'augmenter la performance de l'organisation quant à sa réponse aux multiples enjeux.

I. Le secteur bancaire : marché, mutations et perspectives

I.1. Un secteur économique consolidé et universalisé

Les banques sont des établissements financiers dont les activités se répartissent entre opérations de crédits, opérations financières et mise à disposition de moyens de paiement.

I.1.1. Le secteur en France en 2018

Le marché du secteur bancaire français compte à ce jour 347 banques dont 6 groupes bancaires nés de divers regroupements et d'une forte dynamique de marché[1] :

- BNP Paribas,
- Le Groupe Banque Populaire – Caisse d'Epargne,
- Le Groupe Crédit Mutuel – CIC,
- Le Groupe Société Générale,
- Le Groupe Crédit Agricole,
- La Banque Postale.

En 2018 en France :

- Ces groupes gèrent près de 80% des comptes courants et disposent pour la plupart de réseaux internationalisés.[2]
- Les banques contribuent à hauteur de 2,1% de la valeur ajoutée totale du pays et sont le premier employeur national avec 2% des salariés.
- Quatre banques françaises sont parmi les neuf premières en Europe et deux d'entre elles sont parmi les dix plus importantes banques au monde.
- Elles contribuent au financement de l'économie via l'octroi de prêts immobiliers à hauteur de 967 milliards d'euros d'encours à fin avril 2018 selon la Banque de France.
- L'épargne réglementée s'élève à environ 386 milliards d'euros et l'encours des contrats d'assurance vie s'élève à 1697 milliards d'euros.
- Les moyens de paiement se diversifient et si 58480 distributeurs maillent le territoire, les français utilisent de plus en plus les nouveaux moyens provenant de la transformation numérique de l'économie : Accès via tablette et smartphone et paiement sans contact se développent rapidement.

I.1.2. Une inflation règlementaire en réponse aux crises financières, aux enjeux internationaux et aux mutations économiques

Piliers de l'économie mondiale, les groupes bancaires doivent assurer en permanence divers services de mise à disposition de moyens de paiement et de crédit dans des logiques de concurrences parfaites.

---

[1] Observatoire des métiers de la banque | les acteurs du système bancaire observatoire-metiers-banque.fr | consulté en octobre 2018.

[2] Fédération française des banques | Faits et Chiffres N°01 - le secteur bancaire français fbf.fr | publié en juillet 2018 | consulté en octobre 2018

Ainsi, les législateurs et exécutifs internationaux ont mandaté des régulateurs quant au suivi de nombreuses règlementations.

Parmi les nombreuses règlementations nous pouvons principalement citer :

- Les accords de Bâle précisant les modalités de fonctionnement intra et interbancaires,
- Les sanctions internationales obligeant les entreprises du métier à filtrer les flux financiers en provenance et à destination de pays sous embargos,
- La lutte contre le blanchiment et financement du terrorisme précisant entre autre comment les banques doivent s'assurer de l'identité de leurs clients,
- Les règlements et lois sur la sécurisation des nouveaux moyens de paiement en réponse à la croissance de leur piratage.

D'autres lois issues d'initiatives politiques ont été promulguées. i.e. les lois Hamon ou Eckert

Fait marquant de 2014 : BNP Paribas a été condamné à verser une amende de 6,5 milliards d'euros à l'Etat américain suite à des transactions effectuées en dollars avec des pays sous embargo américain. Le groupe n'avait pas respecté des lois américaines.

I.1.3. Le modèle de banque universelle

D'après leur document de référence, les principaux groupes bancaires ont pour caractéristiques communes, de regrouper leurs métiers en quatre familles d'activités :

- La banque de détail :

Destinée au grand public, la banque de détail a pour rôle de fournir des services de crédit, d'épargne et de moyens de paiement aux particuliers et aux petites et moyennes entreprises.

- La gestion d'actifs :

Fournissant des services de placement et d'épargne structurés selon divers critères, l'activité d'Asset Management permet aux clients de faire fructifier leur patrimoine financier.

- La banque de financement et d'investissement :

Destinées aux grandes clientèles, l'activité de BFI se distingue dans les projets financiers importants impliquant les grandes entreprises et les Etats.

- Les métiers spécialisés :

Regroupant divers métiers spécialisés ou contingents à l'activité bancaire, ils regroupent entre autre, les activités de crédit-bail ou d'affacturage.

Ces quatre familles d'activités constitient le modèle de banque universelle associant différents types de métiers autour de la bancassurance et de la finance.

I.1.4. Les Fintechs

Elles s'appellent Boursorama, B For Bank, ING Direct, Hello Bank, N26, Revolut ou My French Bank. Ces banques d'un nouveau genre, 100% digital, sont issues de la fusion de transformation numérique de l'économie et du métier bancaire. En mettant le client au cœur de l'expérience utilisateur, il devient autonome et peut réaliser l'ensemble de ses opérations en ligne depuis tout type de canal électronique. Leur attrait est symbolisé par une grille tarifaire particulièrement attrayante.[3]

[3] Le patron d'Orange Bank nous raconte les néobanques | Guerric Poncet
lepoint.fr | publié en aout 2018 | consulté en octobre 2018

Les Fintechs sont les nouveaux relais de croissance du secteur bancaire. Ainsi de nombreuses banques ont massivement investi en Recherche et Développement dans la « Blockchain » et la fluidification des moyens de paiement.

I.2. Contextualisation du Groupe Crédit Agricole – un équilibre entre continuité et mutations

I.2.1. Introduction

Communément appelée la banque verte du fait de sa proximité historique avec les métiers agricoles, le Crédit Agricole est le dixième groupe bancaire mondial en 2018. Son modèle mutualiste érigé en caisses locales, régionales puis nationale offre un modèle de stabilité et universel au service du client.

I.2.2. Historique

I.2.2.1. La naissance du Crédit Agricole Mutuel

A la fin du 19ème siècle, le secteur agricole peine à trouver des financements nécessaires à son développement. C'est sur une initiative de Louis Milcent en 1885 qu'est créée la Société de crédit agricole de l'arrondissement de Poligny dans le Jura[4]. En réunissant les agriculteurs, le Crédit Agricole Mutuel devient une alternative aux modes de de financement classiques de l'agriculture de la fin du 19ème siècle.

a. Siège de la Société de Crédit Agricole Mutuel du Jura fondée en 1885. Bâtiment hébergeant désormais la Fondation Maison de Salins.

Photographie utilisée avec l'aimable autorisation de la Fondation de la Maison de Salins.

I.2.2.2. L'essor des caisses locales et régionales

En 1894, Jules Méline, homme politique et défenseur de l'agriculture est à l'origine de la loi du 5 novembre 1894 favorisant la création de sociétés de crédit agricole[5] : la première caisse

[4] Groupe Crédit Agricole | Histoire du Groupe Crédit Agricole
credit-agricole.com | consulté en octobre 2018

[5] Ministère de l'agriculture | les textes fondateurs du monde agricole - Loi du 5 novembre 1894 relative à la création de société de crédit agricole.
agriculture.gouv.fr | consulté en Octobre 2018

locale est créée. De nombreuses caisses locales composant le premier étage de la pyramide, seront réparties sur le territoire. Afin de compenser les carences en capitaux nécessaires aux financements, l'Etat impose en 1897 à la Banque de France, un apport de 40 millions de francs or et une redevance annuelle de 2 millions de francs au Crédit Agricole Mutuel.

La loi du 31 mars 1899 autorise le regroupement des caisses locales en Caisses Régionales du Crédit Agricole Mutuel [6]. C'est le second étage de la pyramide. Le succès du dispositif de financement de l'Etat a bénéficié en 1910 à 96 Caisses Régionales.
A la veille de la première guerre mondiale, chaque département bénéficie d'au moins une caisse régionale.

Les prêts à court terme représentent la majeure partie des activités des caisses et bien que la collecte progresse, les trois quarts des ressources sont assurés par l'Etat.

Pendant la première guerre mondiale, les femmes remplacent les hommes dans les champs. Ensuite le Crédit Agricole est sollicité pour financer la mise en valeur des terres abandonnées et pour rétablir les exploitations proches des lignes de front. Les mutilés et les victimes civiles de la guerre bénéficient de prêts avantageux finançant l'acquisition, l'aménagement ou la reconstitution de petites propriétés rurales.

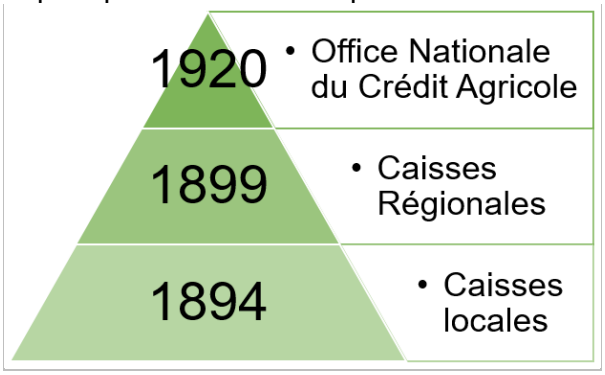
I.2.2.3. La C.N.C.A. troisième niveau pyramidal et la F.N.C.A.

Afin de donner plus d'autonomie à l'institution et pour permettre une compensation centrale entre les Caisses Régionales, est créé en 1920 l'Office National du Crédit Agricole qui deviendra en 1926 la Caisse Nationale du Crédit Agricole (C.N.C.A.).

.

C'est la structure pyramidale du Crédit Agricole connue jusqu'aujourd'hui (sous d'autres appellations).

Le premier Directeur Général de la C.N.C.A. fut Louis Tardy, qui, en occupant plusieurs fonctions au sein du groupe, y aura totalisé 60 ans.



b. Structure Pyramidale du Crédit Agricole

Dans les années 30, La C.N.C.A. aide les caisses locales et régionales les plus exposées aux répercussions du krach de 1929.

Puis, pendant la seconde guerre mondiale, le régime de Vichy mit sous tutelle le Crédit Agricole. Pendant cette période, Louise Tallerie, première Directrice de caisse régionale sauva des archives durant la phase d'exode.

Le Crédit Agricole participe activement à la reconstruction du pays de l'après-guerre et à la mécanisation de l'agriculture, par l'ouverture de nombreux bureaux maillant le territoire français, émettant des bons à 5 ans et des obligations à long terme. Il s'émancipe progressivement des financements de l'Etat puis s'autofinance à partir de 1963.

L'année 1947 est marquée par la création de la Fédération Nationale du Crédit Agricole (F.N.C.A.) à la Rochelle. La F.N.C.A. est l'organe parlementaire du Groupe Crédit Agricole.

---

[6] Ibid.- loi du 31 Mars 1899 ayant pour but l'institution des Caisses Régionales de Crédit Agricole Mutuel et les encouragements à leur donner ainsi qu'aux sociétés et aux banques locales de crédit agricole mutuel
agriculture.gouv.fr | consulté en Octobre 2018

Elle permet d'échanger entre les différentes entités du Groupe et de fixer ses orientations politiques.

I.2.2.4. La diversification des activités et des enseignes

A compter des années 60, le Groupe diversifie ses activités au-delà du secteur agricole. Afin de développer son activité, la banque obtient son autonomie financière en 1966. Elle étend progressivement son champ d'activité aux zones rurales et aux industries agro-alimentaires. En 1971, elle peut émettre des financements à destination des PME et des PMI. Puis, en 1981 et 1982, le groupe peut financer les entreprises et les ménages. En 1985, les commerces de détail et l'hôtellerie-restauration peuvent être financés. Avec le financement des grandes entreprises en 1991 s'achève l'extension de son champ de compétence.

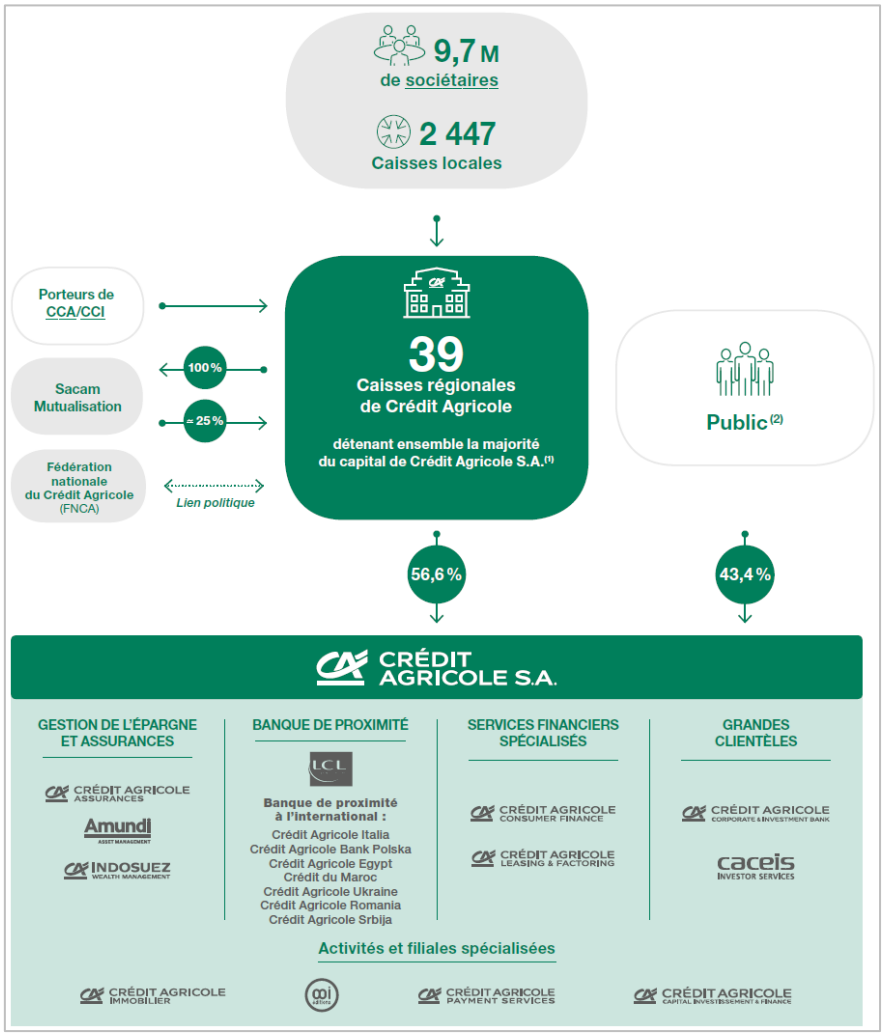
Afin de répondre à une diversité de demandes croissantes du marché, de nombreuses filiales spécialisées dans divers métiers sont créées ou achetées via des plans de fusion-acquisition :

- En 1962 est créée SOFIDECA afin de prendre des participations dans les entreprises agroalimentaires
- En 1967 est créée SEGESPAR pour les prises de participations du Crédit Agricole
- Les sociétés Unimat et Unicomi sont créées en 1969 afin d'entrer sur les marchés du crédit-bail mobilier et immobilier
- Unicrédit est créée en 1971 pour les crédits aux industries agro-alimentaires
- En 1972, sont créées les sociétés d'assurance-vie SORAVIE du CEDI ancêtre de Crédit Agricole Payment Services (CAPS)
- Tournant historique : en 1988, les Caisses Régionales rachètent la caisse nationale à l'Etat qui devient une Société Anonyme majoritairement détenue par les Caisses Régionales.
- Initialement partenaire de Groupama dans le secteur de l'assurance, le Groupe créé en 1986 Predica puis en 1990 Pacifica qui seront regroupées sous l'entité Crédit Agricole Assurances, leader de la bancassurance.
- En 1996, est rachetée la banque d'affaire historique Indosuez dont les origines remontent à 1875.
- En 2000 est rachetée Sofinco puis en 2003 Finaref. Ces entités seront fusionnées en 2010 sous l'entité Crédit Agricole Consumer Finance.
- En 2001, le Groupe entre en bourse.
- En 2003 le Crédit Lyonnais entre également dans le Groupe.
  Le Crédit Lyonnais récemment devenu le LCL est une banque créée en 1863 par un groupe d'hommes d'affaires spécialisés dans le commerce de la soie. Développé d'abord dans la région lyonnaise et dans l'est de la France, le Crédit Lyonnais maille tout le territoire français et possède des bureaux à l'international le rendant première banque du monde par la taille de son bilan en 1914.
- En 2004 sont fusionnées les activités de banque de financement et d'investissement de Crédit Agricole Indosuez et du Crédit Lyonnais dans l'entité Calyon qui deviendra Crédit Agricole Corporate & Investment Bank en 2010.
- Les crises financières des subprimes (prêts hypothécaires à risque) et de la dette souveraine de la période 2007 – 2012 entraineront le recentrage de l'activité par la vente ou la cessions d'actifs d'un certain nombre de filiales internationales ou spécialisées dans le courtage dont Emporiki, Bankinter, Cheuvreux, CLSA et Newedge.

- 2010 est marqué par la création d'Amundi, filiale spécialisée dans la gestion d'actifs. C'est une entreprise issue de la fusion de Crédit Agricole Asset Management et Société Générale Asset Management qui sera majoritairement détenue par le Groupe Crédit Agricole. Le Groupe Société Générale cèdera ses actifs courant 2015. Puis courant 2017, Pioneer Investments sera absorbé par Amundi.
- En 2018 Indosuez Wealth Management lance Azqore, qui est la nouvelle marque de Crédit Agricole Private Banking Services (CA-PBS)

I.3. Le groupe en 2018

Le Groupe compte désormais 139 000 collaborateurs au service de 52 millions de clients dans 49 pays.



c. Schéma présentant la structure globale du Groupe Crédit Agricole
Illustration provenant du document de référence 2017 du Groupe Crédit Agricole

Sa structure est composée par un actionnariat détenu en majorité par les Caisses Régionales de Crédit Agricole et en minorité par le public composé d'investisseurs, de salariés et d'actionnaires individuels. Au sein du Groupe Crédit Agricole, Crédit Agricole S.A. est la banque centrale et l'organe qui garantit l'unité financière du Groupe et veille au bon fonctionnement du réseau Crédit Agricole. Il coordonne les stratégies des filiales spécialisées du Groupe en France et à l'international. En 2001, l'entrée en Bourse de Crédit Agricole S.A. marquait la volonté des Caisses Régionales d'étendre leur développement en France et en Europe. Ses activités sont organisées en 4 lignes métiers comprenant diverses enseignes.

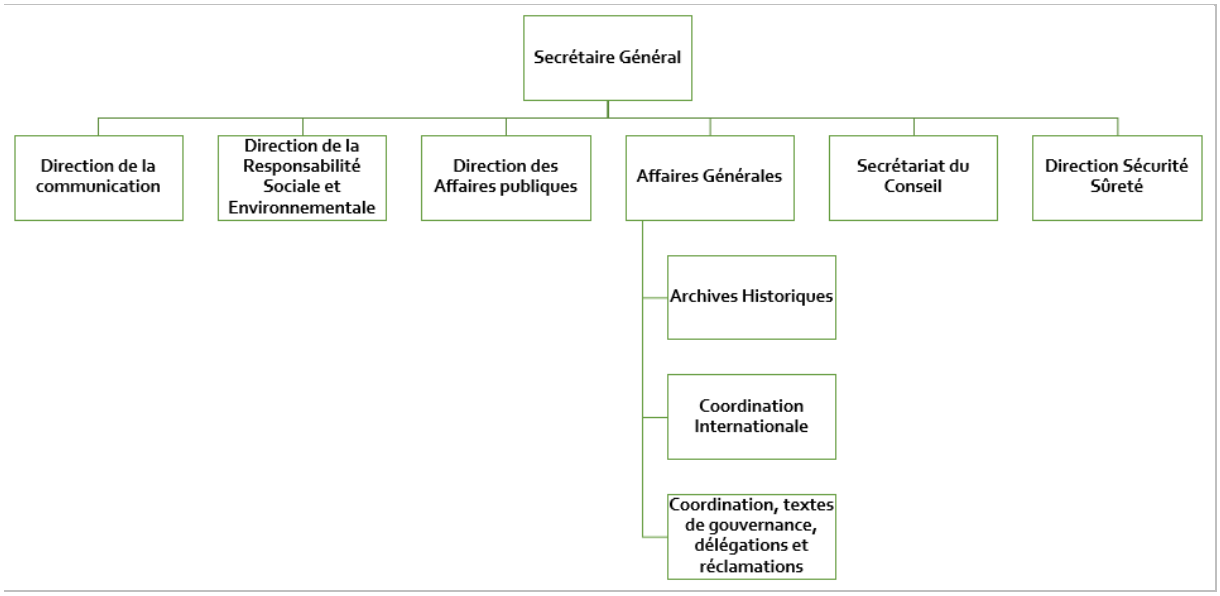
La Fédération Nationale de Crédit Agricole (FNCA) est l'organe de réflexion et de représentation des Caisses Régionales, lieu où sont débattues les grandes orientations du Groupe.

II. Problématique : améliorer l'accès à la documentation, à l'information et aux connaissances.

II.1. Environnement de la mission : les Affaires Générales au sein de Crédit Agricole Société Anonyme

Banque centrale du Groupe Crédit Agricole, Crédit Agricole S.A. est la holding qui veille à la cohérence du fonctionnement du groupe et à son unité financière.

Son organigramme est composé de l'ensemble des métiers du groupe. Puis, afin d'assister la Direction Générale de Crédit Agricole S.A. dans ses actions, le Secrétariat Général œuvre auprès du Directeur Général dans les domaines décrits ci-dessous :



d. Organisation du Secrétariat Général

Les Affaires Générales ont pour rôle :

- La gestion des archives historiques du groupe,
- La coordination internationale,
- La gestion et le suivi des textes de gouvernance et de la chaîne délégataire.
  Ce service gère aussi les réclamations adressées à la holding Crédit Agricole S.A..
  Les textes sont regroupés dans un corpus comprenant des procédures et des informations générales destinées au Groupe.
  Situés sur l'intranet des Affaires Générales, ces documents structurés sont consultés par plusieurs milliers d'utilisateurs du groupe.
  Cette activité est confiée au responsable de ce service qui veille à la cohérence du corpus et à la publication des documents.

II.2. L'intranet des Affaires Générales - description

II.2.1. Rôle et usage de l'intranet des Affaires Générales

L'intranet des Affaires Générales est dédié à la publication de diverses informations transverses stratégiques et opérationnelles à destination du Groupe.

Ce sont majoritairement :

- Des procédures liées aux différentes activités, métiers et fonction du Groupe,
- Des textes de gouvernance concernant l'organisation de l'entreprise,
- Et des circulaires destinées spécifiquement aux Caisses Régionales du Crédit Agricole. Ces circulaires reprennent le contenu normatif des procédures, mais aussi les informations financières et marketing essentielles au fonctionnement du réseau. Ces circulaires se nomment des « lettres jaunes » car historiquement imprimées sur du papier de couleur jaune afin d'être identifiable avant publication dans des formats numériques.

Ces documents sont pour partie, les déclinaisons opérationnelles de la législation, ils ont aussi pour sujet des analyses sectorielles ou les produits commercialisés par le Groupe.

Les documents sont normalisés et respectent des gabarits.

L'intranet est basé sur SharePoint 2013 ; il héberge des documents PDF et bureautiques comportant plusieurs métadonnées métiers.

Son audience est estimée à environ 200.000 visites annuelles et il comprend environ 10.000 documents.

Cet intranet contient l'historique des informations publiées depuis 2002. Auparavant les documents étaient communiqués au format papier aux parties prenantes. Les versions papier sont aujourd'hui disponibles aux archives historiques.

L'usage de cet intranet permet de diffuser des informations immédiatement applicables à leur publication.

Toutes les directions des entités du Groupe et la Fédération Nationale peuvent solliciter les Affaires Générales, à des fins de publications dont le processus de conception et de validation est décrit dans le schéma suivant.

II.2.2. Processus de demande de publication

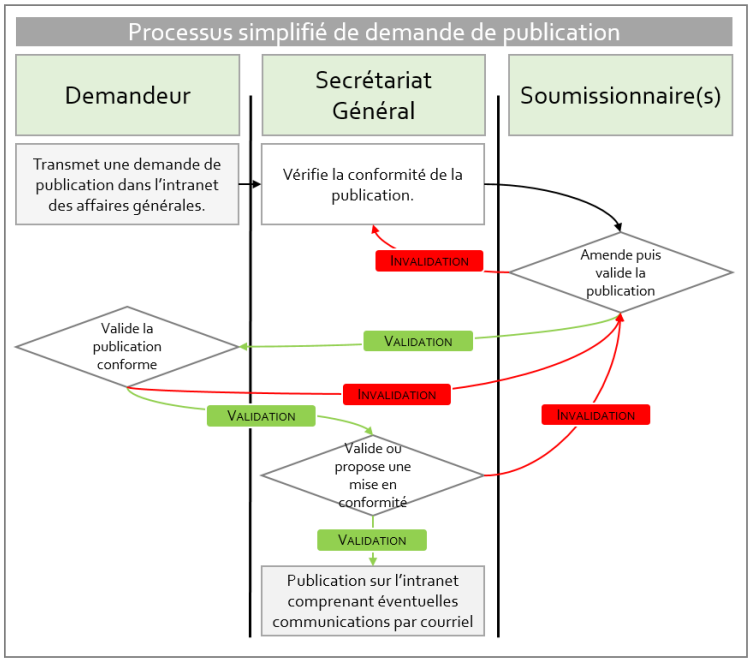
Une demande de publication est un processus complexe impliquant plusieurs intervenants. Il a été modélisé en le simplifiant et en rendant génériques les étapes et les acteurs.

- **Le demandeur** est l'acteur qui initie une demande de publication de texte.

- **Le secrétariat général** est l'acteur en charge de veiller à la conformité de du texte et à sa publication.

- **Le(s) soumissionnaire(s)** sont le(s) acteur(s) validant les documents.

Dans certains cas, le soumissionnaire et le demandeur peuvent être le même acteur.



e. Processus simplifié de demande de publication sur l'intranet des Affaires Générales

II.3. Des demandes d'améliorations de l'intranet

II.3.1. Enquête de satisfaction

Afin d'évaluer la qualité du service rendu, une enquête de satisfaction a été menée auprès des utilisateurs de l'intranet courant 2016. Elle a mis en évidence une satisfaction globale des utilisateurs via une note moyenne de 7,3/10. D'autre part, ils ont manifesté des souhaits d'amélioration quant à l'accès aux informations et aux documents.

II.3.2. Des améliorations en mode projet

Dans la continuité de l'enquête de satisfaction, un projet ayant pour objectif la mise à disposition d'une nouvelle offre d'accès à l'information basé sur la compréhension des besoins des utilisateurs, a été édifié pendant l'été 2018.

Il a été décidé de comprendre ce besoin via des analyses statistiques car :

- Vu que le projet eut lieu durant la période estivale, les utilisateurs étaient fréquemment en congés. En conséquence de quoi, il était difficile de les contacter dans l'objectif d'obtenir leurs avis.
- L'enquête de satisfaction ayant déjà été réalisée, il n'était pas concevable d'en refaire une nouvelle
- Au regard de la dispersion géographique des utilisateurs et dans les filiales, il aurait été difficile d'obtenir des avis représentatifs provenant de diverses franges de la population.

Les analyses statistiques ont porté sur :

- L'analyse de l'audience des utilisateurs sur l'intranet,
- Puis sur l'analyse du corpus des Affaires Générales via le text-mining.

II.4. Méthodologies mises en œuvre - conduite de projet

La réussite d'un projet dépend de nombreux paramètres dont l'organisation qui lui est accordée. En associant différentes méthodes, il devient possible de donner une visibilité rassurant les commanditaires et conduisant à des livraisons sans pression.

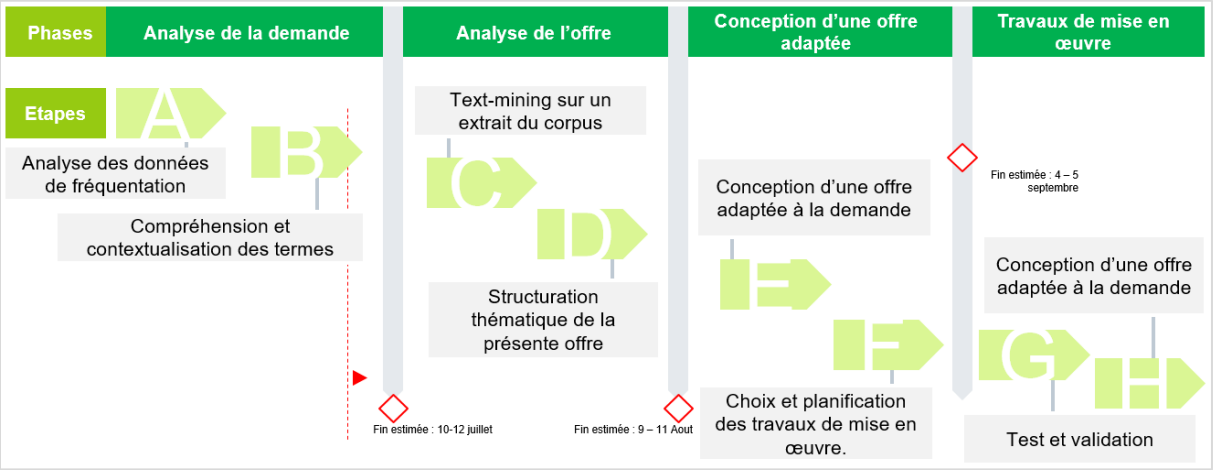
Le projet global a nécessité plusieurs méthodes de conduite de projet.

Par l'association de diverses techniques issues du PMBOK (Project Management Body of Knowledge), d'Agile, de diverses méthodes de recherches et de la psychologie professionnelle, il est plus aisé d'atteindre l'objectif défini dans le projet.

II.4.1. Planification via un macro-planning de type Gantt

Un projet comporte différentes phases chronologiques pendant lesquelles des actions et des tâches se succèdent.[7]

En ce qui concerne le projet de compréhension des besoins des utilisateurs de l'intranet, nous avons établi le planning ci-dessous découpant le projet en phases dont la fin est jalonnée par une présentation validant le travail effectué.



f. Illustration présentant les phases du projet

Nous avons structuré le projet en quatre phases comportant chacune deux étapes :

- L'analyse de la demande : phase d'analyse de l'audience de l'intranet afin d'en dégager des tendances de fréquentation et de recherches.
- L'analyse de l'offre : phase associant techniques de text-mining et de compréhension du corpus par le vocabulaire utilisé.
- La conception d'une nouvelle offre adaptée : alignement de l'offre (corpus) sur la demande (audience) via la mise en place de nouveaux accès à l'information.
- La mise en œuvre des travaux : évaluation des durées et coûts du projet puis lancement.

II.4.2. Méthodes agiles : alignement du projet d'après le besoin du commanditaire

Les courtes durées de projets, les environnement complexes et l'évolutivité de la demande nécessitent une adaptation permanente des méthodes de travail en vue de transmettre les livrables au plus proche du besoin dans des délais toujours plus courts.[8]

Les méthodes agiles permettent de se concentrer sur la réponse au besoin du client. Disposant de nombreuses pratiques, il n'est pas nécessaire de faire usage de l'ensemble des méthodes. Le choix de quelques pratiques aide à l'avancement d'un projet.

---

[7] Conduite de projet informatiques. Développement, analyse et pilotage (4e édition) | Brice Arnaud Guérin
ENI | Parution : aout 2018 | ISBN : 9782409014635
Livre numérique consulté sur eni-training.com en octobre 2018
Chapitre : « *La planification et le chiffrage* » > « *La planification* » > « *1. Les éléments d'un planning* »

[8] La méthode Agile à grande échelle | Andy Noble, Darrell K. Rigby, Jeff Sutherland
hbrfrance.fr | publié le 13 septembre 2018

La durée particulièrement courte du projet a nécessité des ajustements réguliers. En réalisant des sprints dont les livrables étaient présentés en fins de phases, le projet était découpé de façon rationnelle. Des points intermédiaires, rapides et réguliers permettant d'ajuster les livrables au plus proche du besoin.

II.4.3. Synchronisation avec son interlocuteur

Un projet implique des phases de rencontres avec le commanditaire lors de réunions, de déjeuners ou pendant des points informels. Ces échanges d'informations sont à l'origine des actions menant à la réussite du projet. [9]

La synchronisation, élément important dans la réussite d'un entretien ou d'un échange consiste en l'écoute active, la reformulation et la prise en compte des informations de l'interlocuteur.

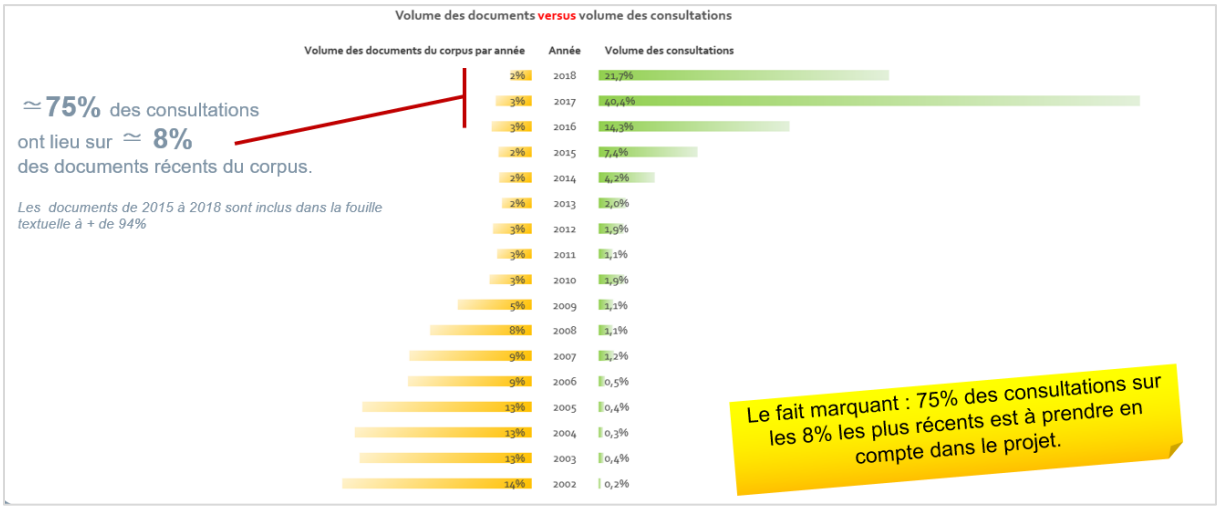
Au regard de la complexité des enjeux du projet, il a été nécessaire d'adapter en permanence l'écoute afin de construire un échange de qualité.

[9] Manager un équipe projet | Pieric Couteaud Horrut
Support du cours du 8 février 2018

II.4.4. Visibilité : partager et comprendre des données complexes via des infographies et Datavisualisation

Plusieurs livrables contenaient de nombreuses informations chiffrées complexes. A partir de données statistiques, de nombreuses données chiffrées sont extraites. La représentation via la datavisualisation et sous la forme d'infographie permet de mieux comprendre des informations en vue de prendre des décisions.[10]

En ce qui concerne le projet de compréhension des besoins des utilisateurs, les données ont été systématiquement présentées sous la forme de graphiques ou via la datavisualisation afin d'en dégager rapidement les enjeux.



g. Datavisualisation : "Butterfly Chart" exprimant les volumes de consultations par année versus les volumes de documents

<u>Commentaires sur le schéma</u> :

Les documents du corpus sont répartis dans des rubriques représentant les années de publication. Pour chaque année, la partie gauche du schéma représente le volume de documents généré, tandis que la partie droite illustre son audience.

L'illustration met en évidence que 75% des consultations des intranautes sont réalisées sur les 8% de documents les plus récents.

C'est un fait marquant ayant impacté le projet.

La datavisualisation met ce fait en perspective ; ce qui permet d'identifier et partager des enjeux.

[10] DataViz Quels outils pour quelles datavisualisations ? | Serge Courrier | Page 4 Slideshare.net | mise à jour publiée en septembre 2017 | consulté en Octobre 2018

II.4.5. Recherches : trouver la littérature de qualité et des pratiques informatiques

II.4.5.1. Littérature en rapport avec la textométrie (text-mininig)

Le travail de chef de projet en ingénierie documentaire et gestion des connaissances nécessite de fréquemment rechercher l'information essentielle à une activité ou pour prendre des décisions.

Dans le cadre du projet, la recherche documentaire et informationnelle a été un point essentiel à sa réussite.

Afin d'analyser le texte du corpus comprenant près de 10000 documents il a été nécessaire de s'initier à textométrie. Plusieurs recherches via le portail des Sciences Humaines & Sociales CAIRN, le moteur de recherche de Google et les sites comprenant des mémoires et thèses ont contribué à identifier et comprendre des informations clés sur la lexicométrie.

Les recherches ont permis de connaitre les sites internet des JADT – Journées internationales d'Analyse statistique des Données Textuelles qui comprennent des retours d'expériences et résultats de recherches résumés dont il est intéressant de s'inspirer.[11]

II.4.5.2. Comprendre le jargon de la banque et de la finance

Chaque domaine d'activité économique dispose de son langage métier. le domaine bancaire, relativement complexe et réglementé est caractéristique d'une terminologie volumineuse composée d'acronymes et de termes dont l'usage et la définition peuvent changer dans le temps.

Comprendre le projet a nécessité des recherches avancées sur les bibliothèques numériques et sur internet. Il a été aussi nécessaire de consulter de professionnels de la Finance.

II.4.5.3. Recherches de connaissances en scripting

Travailler sur un corpus de taille importante de documents nécessite de convertir, nettoyer et fusionner plusieurs milliers de textes. Des recherches avancées sur des forums de développeurs tels que Stackoverflow.com ont donné l'accès à des connaissances en scripting et en SQL ayant permis d'industrialiser la préparation du corpus.

La préparation du corpus implique par exemple, la suppression de certains caractères spéciaux ne devant être dédiés qu'au balisage du texte pour les logiciels de text-mining. L'aide des forums a permis de trouver des scripts permettant d'automatiser la suppression de ces caractères spéciaux.

II.4.6. Stimulation de la créativité

La complexité technique du projet a souvent nécessité la stimulation de la créativité.

Divers travaux de recherches ont démontré l'importance de la déconnexion aux outils numériques (dont les notifications des téléphones intelligents) dans la stimulation des processus liés à la concentration.[12]

---

[11] Analyse de données textuelles
fr.wikipedia.org | consulté en novembre 2018

[12] "Silence Your Phones": Smartphone Notifications Increase Inattention and Hyperactivity Symptoms
Kostadin Kushlev, Jason Proulx, Elizabeth W. Dunn
DOI : http://dx.doi.org/10.1145/2858036.2858359 | publié en mai 2016 – consulté en octobre 2018

Aussi, passer un temps de pause dans des lieux reposants ou ayant de la verdure contribuent à stimuler la créativité.[13]

Le siège de Crédit Agricole S.A. est situé sur le campus Evergreen pensé comme un village[14]. Il dispose d'un parc de 4 hectares composé de plusieurs bassins et rivières. Y sont présentes plus de 90 espèces végétales différentes et plusieurs espèces animales.

Quelques pauses dans le parc d'Evergreen stimulaient la créativité. L'esprit chargé de nouvelles idées de retour au bureau, il devient plus facile de les mettre à contribution dans un objectif d'avancement du projet.

II.4.7. Conclusion intermédiaire – des méthodologies au service de la performance projet

La conclusion intermédiaire est une étape analysant les faits marquants d'une phase du projet. En résumant les principaux points, elle permet de mettre en perspectives des éléments déterminants à la poursuite du projet.

Un projet réussi nécessite une forte implication et de l'organisation. Les méthodes de travail décrites dans le cadre de ce projet ont été des facteurs de réussite. Il est important de prendre en considération cette dimension donnant un nouveau relief aux métiers de l'info-doc et de l'info-connaissance.

L'association équilibrée des fondamentaux du management de projets à la recherche documentaire et à quelques points de psychologie professionnelle, est un point essentiel quant à l'atteinte des objectifs.

[13] Regarder la nature rend plus productif | Nicole Torres
hbrfrance.fr | publié le 10 mars 2018 | consulté en octobre 2018

[14] Visite de l'éco-campus Evergreen, nouveau siège du Crédit Agricole S.A. | Fabrice Mazoir
Blog-emploi.com | publié en juillet 2014 | consulté en octobre 2018

II.5. Comprendre le besoin en analysant l'audience du site et le contenu du corpus

En entreprise, la transformation numérique des activités a entrainé une numérisation des données analogiques, puis la génération de nouvelles données nativement numériques.

De façon similaire, l'intranet des Affaires Générales a permis le déplacement d'une partie du corpus papier numérisé sur l'outil de collaboration Sharepoint. Les nouveaux documents sont nativement numériques. La consultation se fait via un navigateur internet.

La structure numérique des documents et de leur hébergement permet d'analyser :

- L'audience de l'intranet,
- Le contenu du corpus.
  De nouvelles perspectives d'analyses des fréquentations et de compréhension du corpus via le text-mining (connu sous les noms de textométrie ou de fouille textuelle) ont été exploitées.

La puissance de calcul grandissante des ordinateurs associée aux nombreuses initiatives professionnelles et universitaires offrent de nouvelles possibilités d'analyses et de recherches statistiques, en vue de mieux comprendre les comportements des utilisateurs et le contenu des corpus.

II.5.1. Structure du site de l'intranet des Affaires Générales

L'intranet des affaires générale est basé sur SharePoint 2013. Il héberge le corpus des Affaires Générales comprenant des procédures, des circulaires et des analyses sectorielles.

En majorité, ces documents font office de directives applicables à leur publication.

L'intranet comporte :

- Plusieurs rubriques documentaires; **chacune correspondant à un type de document.**
- Plusieurs sous-rubriques : chacune d'elle correspondant à une année de publication située entre 2002 et 2018.

Il y a 10000 documents environ, suivant diverses normes de présentation et sont à 90% au format PDF. Plusieurs métadonnées plus ou moins exploitées sont associées aux documents.

Rubriques documentaires	Description
Notes d'organisation et notes de nomination	Description des mises jour organisationnelles de la banque et publication des nominations à des fonctions clés.
Notes de procédures et de fonctionnement	Ensemble de normes groupes qui sont pour partie les déclinaisons opérationnelles du code monétaire et financier.
Lettres Jaunes	Circulaires reprenant les notes de procédures mais aussi des analyses sectorielles et des descriptifs de produits commercialisés par le groupe.

1. Description des rubriques du site

II.5.2. Analyse de l'audience

Dans la continuité de l'enquête de satisfaction de 2016, une analyse complète de l'audience a été réalisée.

L'audience de l'intranet SharePoint des Affaires Générales est mesurée grâce un fournisseur tierce offrant des analyses clés en main et la possibilité d'extraire des données brutes.

Ainsi, il a été possible d'ériger plusieurs statistiques d'usages ayant permis de comprendre les intérêts des utilisateurs.

Les statistiques ont été réalisées sur la base de 2 années glissantes à des fins de comparaison : 1er juillet 2017 au 30 juin 2018 versus 1er juillet 2016 au 30 juin 2017

Fréquentation de l'intranet

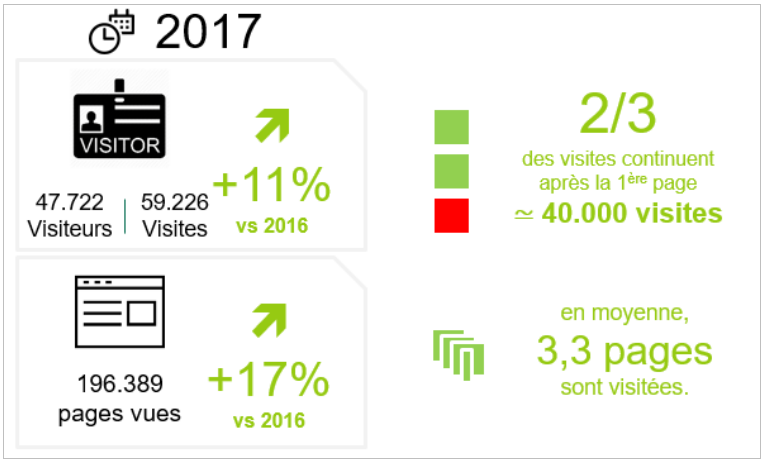
La fréquentation de l'intranet donne une tendance sur l'usage global du site. Elle permet de comprendre l'intérêt des intranautes pour le site, leur comportement et si l'information est rapidement retrouvée.

Commentaires :

1. Avec une hausse globale de la fréquentation, nous pouvons constater un intérêt grandissant pour le site.

2. 2/3 des visiteurs continuent après la 1ère page. Ce qui peut être interprété par le fait qu'un tiers des utilisateurs trouvent l'information sur la page d'accueil.

3. Avec un nombre moyen par session situé à 3,3 pages, nous pouvons constater que l'ergonomie en 3 clics est suivie.



h. Audience : fréquentation de l'intranet

L'intérêt pour le site est globalement en hausse et son ergonomie permet aux utilisateurs de retrouver facilement les informations recherchées.

Consultation des rubriques

Sachant que trois des six rubriques de l'intranet correspondent à des types de documents, la répartition des consultations permet d'évaluer l'intérêt des utilisateurs par type de document.

Rubrique Consultée	Nombre de chargements	Répartition	ECC
Lettres Jaunes	82082	62%	62%
Notes de procédures et de fonctionnement	35217	26%	88%
Notes d'organisation et notes de nomination	12741	10%	98%
Conseil d'administration	1677	1%	99%
Outils / modèles	897	1%	100%
Recherche Avancée	472	0%	100%

i. Audience : répartition des consultations par rubrique

Note : seules les rubriques de notes et de lettres jaunes contiennent des documents.

Commentaires :

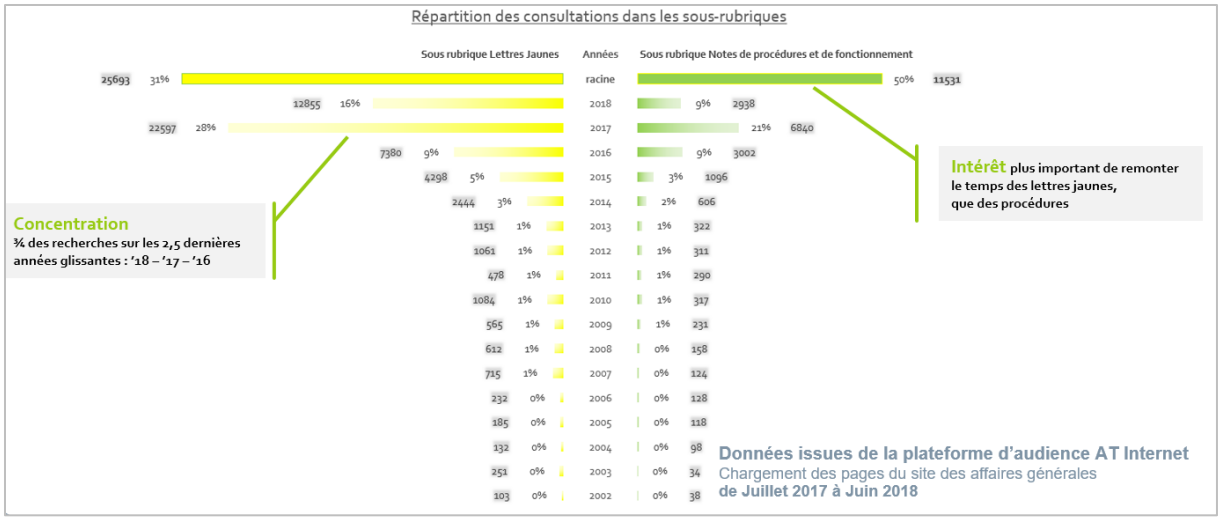
Nous pouvons graphiquement voir que la majeure partie des consultations est réalisée sur les documents. Les lettres jaunes dominent l'audience, tandis que les notes de procédures et de fonctionnement devancent largement les notes d'organisation et de nomination. Nous pouvons aussi voir que l'usage de la recherche avancée reste marginal.

L'effectif cumulé croissant (ECC) des volumes de consultations par rubrique démontre que 98% des chargements ont lieu des documents.

Le volume des consultations des rubriques ne comprenant pas de documents est quant à lui non-significatif.

Consultation des sous-rubriques

L'intérêt majeur étant porté sur les rubriques « lettres jaunes » et « notes de procédures et de fonctionnement », nous avons analysé la fréquentation pour leurs sous-rubriques qui correspondent aux années de publication



j. Audience : répartition des consultations par années pour les rubriques lettres jaunes et notes de procédures et de fonctionnement

<u>Commentaires</u> :

Les sous-rubriques « racine » correspondent à la page d'accueil des rubriques « lettres jaunes » et « notes de procédures et de fonctionnement ». Nous pouvons constater que les utilisateurs trouvent plus facilement les documents recherchés depuis la rubrique « racine » des notes de procédures et de fonctionnement puisqu'ils s'y arrêtent dans 50% des sessions.

Les utilisateurs souhaitant continuer à rechercher des documents par année, se concentrent majoritairement sur les 2,5 dernières années glissantes. Il y a donc un fort intérêt pour les documents récents.

En ce qui concerne les « notes de procédures et de fonctionnement », nous pouvons déduire que la demande de récence correspond à un important souhait de consultation des documents en vigueur. Les notes sont effectivement applicables à leur publication et jusqu'à leur abrogation. Cette déduction a été confirmée par les commanditaires du projet.

II.5.2.1. Usages du moteur de recherche

L'intranet dispose d'un moteur de recherche permettant aux utilisateurs de rechercher des documents d'après des termes saisis dans le champ dédié. Après instruction de la requête, le moteur compare ces termes avec le contenu de l'index afin de restituer des résultats de recherches filtrables selon plusieurs métadonnées.

Près de 30.000 requêtes ont été enregistrées sur la période annuelle Juillet 2017 – Juin 2018.

L'outil d'analyse d'audience ayant enregistré les fréquences d'usage des termes de recherche, il a été possible de dégager des tendances de demandes selon divers critères.

Processus d'analyse de la terminologie employée dans le moteur de recherche

Regroupant pêle-mêle tous les termes de recherches utilisés, les données brutes ont nécessité un traitement adapté en vue d'en dégager un sens.

Le processus suivant décrit les six étapes nécessaires à la compréhension de l'usage des termes utilisés dans le moteur de recherche.

ETAPE	DESCRIPTION
ETAPE 1	**DIFFERENCIATION ENTRE LES TERMES DE RECHERCHE ET LES RECHERCHES DE DOCUMENTS**
Cette première étape consiste en la différenciation entre les termes de recherche et les recherches de documents. *exemple de recherche de termes : « sanctions internationales »* *exemple de recherche de document d'après sa référence : « NP-2018-026 »*	
ETAPE 2	**FILTRAGE DES TERMES DE RECHERCHES DOCUMENTAIRES**
Exclusion des termes de recherche utilisés pour rechercher des documents	
ETAPE 3	**CLASSEMENT DES TERMES DE RECHERCHE DANS L'ORDRE DECROISSANT DE LEUR USAGE**
Afin de savoir quels sont les termes les plus utilisés, les termes de recherches sont classés par ordre décroissant d'usage.	
ETAPE 4	**CONTEXTUALISATION ET RACINISATION**
Afin de regrouper toutes les déclinaisons d'un terme utilisé selon ses orthographes ET ses mots contingents, les 100 premiers termes de l'étape 3 sont analysés. *ex : autour du terme « appétence » ont été regroupés les termes « Appétence au risque » et « appétence aux risques ».*	
ETAPE 5	**RECHERCHE DOCUMENTAIRE AFIN DE COMPRENDRE LE JARGON**
Etape préalable au regroupement thématique des requêtes utilisateur, la recherche documentaire est un passage important dans la compréhension des termes utilisés. Il est nécessaire de rechercher le sens du terme dans le jargon métier et de vérifier que la compréhension déduite correspond à son usage dans le corpus.	
ETAPE 6	**REGROUPEMENT THEMATIQUE**
La dernière étape est le regroupement par thématiques de recherches. Elle permet de mieux comprendre les grandes lignes de ce qui est recherché par les utilisateurs	

2. Audience : les 6 étapes d'analyse des termes employés dans le moteur

Analyse des termes de recherche – racinisation et contextualisation

Les mots utilisés par les intranautes dans le moteur de recherche peuvent prendre différentes formes.

Utilisée dans le cadre de prétraitement de données textuelles, la racinisation consiste en la réduction d'un mot via la suppression des variations du préfixe et du suffixe[15]. Très utilisée dans la fouille textuelle, nous en avons fait usage dans l'analyse de l'audience afin de regrouper l'ensemble des variations orthographiques et grammaticales utilisées. La démarche a aussi permis d'identifier les mots contingents.

Dans le tableau ci-contre, nous pouvons voir les différentes déclinaisons orthographiques et expressives du mot « appétence ».

La différence d'effectif entre le premier résultat et le total démontre l'utilité de la démarche.

Par une utilisation contrôlée de la racine et du contexte d'un terme (c.-à-d. en gardant le sens du terme), on comprend mieux son sens et sa classification dans un thème est facilitée. Grâce à l'opération puis à la recherche de sens effectuée, le terme « appétence aux risques » est classé dans la thématique de recherche « risques ».

Termes recherchés	C
appétence	50
APPETENCE	33
appétence aux risques	13
appetence aux risques	9
appétence au risque	2
matrice appétence	2
Cadre d'appétence au risque	2
matrice d'appétence	2
matrice d'appetence	2
appétence 2015	2
APETENCE	2
appétence externalisation	2
appétence risque	1
appétence aux risques des caisses régionales	1
appétences aux risques	1
APP2TENCE	1
appétence 2016	1
cadre d'appetence	1
appetence aux risque	1
Contrôle du respect de la procédure Groupe pour la gouvernance du dispositif d'appétence au risque	1
déclaration d'appétence	1
Total général	**130**

k. Audience : fréquences d'usage des déclinaisons du terme « appétence »

Afin de faire la différence entre l'effectif du mot initialement instruit dans le moteur de recherche puis décliné via la méthode décrite ci-dessus et l'effectif de ses déclinaisons, nous avons appelé cette seconde valeur : **effectif du terme contextualisé**.

Dans le présent cas :

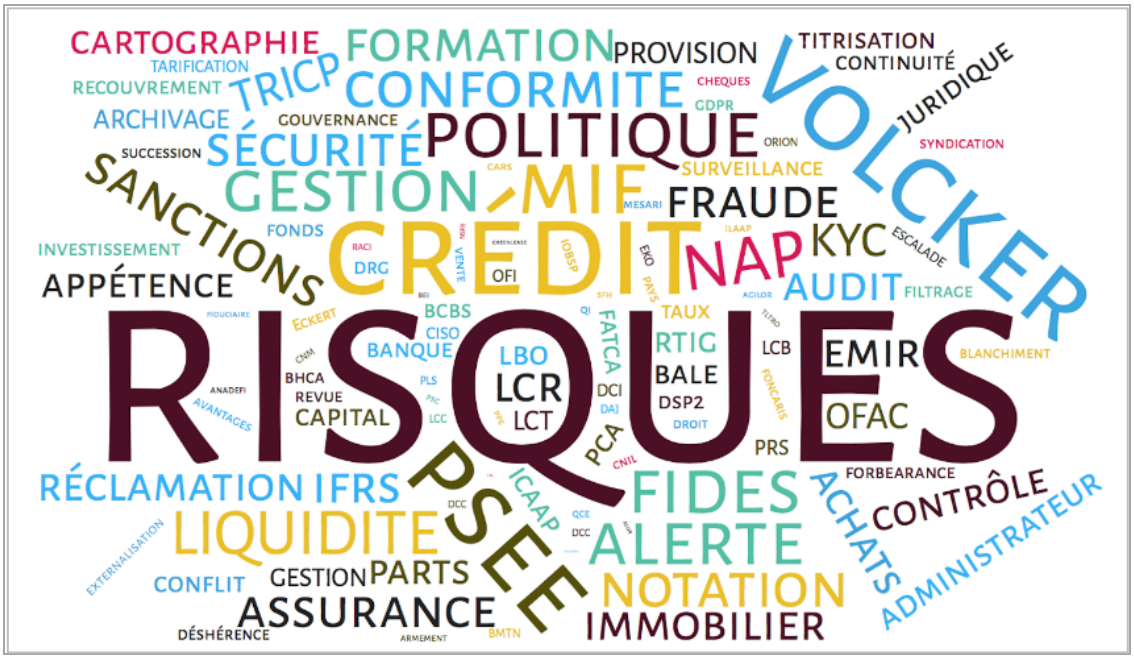
- **L'effectif du terme** est de 50 occurrences,
- **L'effectif du terme contextualisé** est de 130 occurrences.

[15] Modélisation conjointe des thématiques et des opinions | Mohamed Dermouche theses.fr | Page 10 | Thèse soutenue en juin 2015 | consultée en novembre 2018

Nuages de mots

Afin de comprendre visuellement et rapidement les usages contextualisés des mots, nous les avons regroupés dans un nuage de mots.

La taille des mots est en relation avec la valeur de l'effectif du terme contextualisé.



I. Nuage des termes utilisés – taille en fonction de la fréquence d'utilisation

Regroupement en thématiques de recherches

Afin de simplifier l'analyse et la lecture des mots instruits dans le moteur de recherche, les termes contextualisés les plus utilisés ont été regroupés en thématiques de recherche. Ce regroupement résulte d'un travail de compréhension des termes par leur lecture dans le corpus et leur recherche dans la presse spécialisée.

Les thématiques de recherche suivantes ont regroupé les termes contextualisés :

Risques, règlementation, organisation, liquidité, opérations financières, approvisionnement, le client, les produits.

Ces thématiques de recherches associées aux termes utilisés, sont des entrants essentiels à la conception de nouveaux moyens d'accès à l'information.

Répartition des recherches

Afin de comprendre les usages du moteur de recherche, nous avons effectué plusieurs analyses permettant de classer l'usage des termes en trois groupes :

- **Thématiques de recherche :**

Une demande thématique est caractérisée par le souhait de retrouver une ou plusieurs informations sur des sujets précis.

Le travail de thématisation des recherches depuis les termes est assez complexe. Il est décrit dans la suite du mémoire.

Exemple de terme de recherche : « appétence au risque » est un sujet associé à la thématique « Risques ».

- **Recherche de documents :**

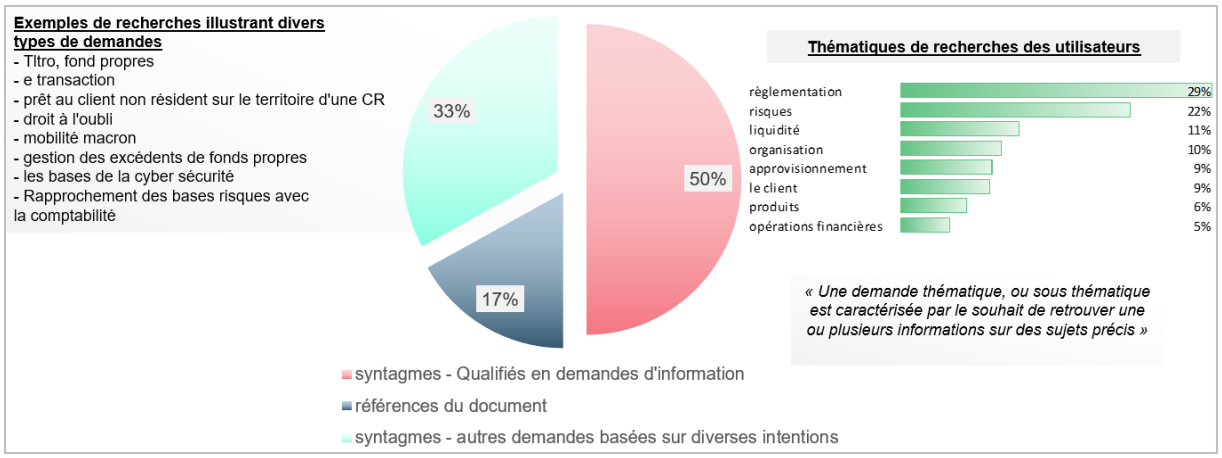
Une recherche de document est caractérisée par la saisie complète ou partielle de la référence de document dans le champ de recherche.

Exemple de terme de recherche : « NP 2018-007 » qui est une référence de note de procédure

- **Autres recherches illustrant divers types de demandes :**

La quantité et le polymorphisme des autres demandes nécessite une étude complémentaire. La déduction n'étant des intentions de recherches n'étant pas évidente, nous en avons déduit qu'il s'agit d'un gisement de demandes en documents, en connaissances et thématiques.

Exemple de terme de recherche : « TLTRO, fonds propres ».



m. Usage du moteur de recherche : répartition par type

Commentaires :

Cette illustration présente la façon dont est utilisé le moteur par les utilisateurs.

De cette analyse, nous pouvons comprendre que les utilisateurs font majoritairement usage de termes groupables en thématiques de recherche dont l'analyse est décrite dans la suite du mémoire. Ils recherchent aussi des documents via leur référence. Dans une moindre mesure, ils utilisent l'intranet telle une base de connaissance, recherchant des cas d'usages précis.

II.5.2.2. Conclusion intermédiaire – Analyse de l'audience

L'analyse de l'audience de la dernière année, réalisée à partir de diverses méthodes et données statistiques, a permis de comprendre ce que les utilisateurs consultent et recherchent.

Nous avons pu en dégager leurs intérêts et des perspectives quant à conception à de nouveaux accès à l'information.

Globalement, nous avons vu que la fréquentation était plus importante sur les lettres jaunes que sur les procédures. Aussi, la majorité des consultations de documents étaient sur les 2,5 dernières années glissantes.

L'usage du moteur de recherche a révélé que les utilisateurs recherchent des documents par leur référence et des informations contenues dans les documents par l'usage d'une terminologie métier. Ils font aussi usage du moteur pour rechercher des connaissances. Ils souhaitent des informations parfois contenues sur d'autres intranets, d'où l'opportunité de les rediriger.

Plusieurs thématiques de recherches ont été identifiées, en relation avec l'actualité du métier, l'organisation de l'entreprise, les nouveaux produits et la règlementation.

II.5.3. Analyse du corpus

II.5.3.1. Pourquoi le text-mining ?

Face à un corpus de taille importante et pour conduire une analyse dans un temps limité, il est difficile de s'en imprégner intégralement par sa lecture. Le corpus des Affaires Générales riche de près de 10000 documents publiés entre 2002 et 2018 comprend une majorité de fichiers PDF correspondant aux circulaires et procédures Groupe faisant office de textes de référence.

Un être humain peut lire un roman à une vitesse moyenne de 300 mots par minute[16]. Si la partie du corpus convertible en fichier texte était un roman, il faudrait y consacrer 21 jours à un rythme de 7 heures quotidiennes afin de parvenir à une lecture complète. Cependant, le contenu de ces documents est en rapport avec des sujets bancaires assez complexes et une lecture linéaire n'en permettrait pas la compréhension. Ce n'est pas un roman.

Cette problématique exprimée, il est nécessaire, afin de comprendre le corpus, d'utiliser des moyens de Traitement Automatique du Langage (T.A.L.) afin d'en dégager un sens, indépendamment de divers biais et sans avis engageant l'analyste. La neutralité conséquente à l'usage de règles statistiques via les outils de TAL permet d'atteindre cet objectif.

La première partie de la mission était consacrée à la compréhension des recherches effectuées par les utilisateurs. Pendant cette première phase de projet, nous avons tenté de comprendre le vocabulaire utilisé par les intranautes afin d'en dégager des thèmes de recherche. Ce premier travail de recherche a permis de comprendre les bases du langage bancaire nécessaire à la seconde phase de projet exploitant des moyens de TAL.

II.5.3.2. Text-mining : quelques faits historiques

Discipline initiée au courant des années 60 par l'analyse du discours, la statistique et le développement de l'informatique, la statistique textuelle commençait à offrir des concepts et des possibilités d'analyse automatisée des textes. Plusieurs travaux de recherches ont contribué à développer de nouveaux concepts et outils informatiques.

Le text-mining, initialement appelé « Statistique textuelle » en France porte plusieurs noms : anciennement lexicométrie, elle se nomme dorénavant fouille textuelle et textométrie.

Le mouvement a été initié par le Professeur Jean-Paul Benzécri, qui a été le précurseur en analyse des données textuelles via la conception de tableaux croisés affichant en colonne des textes, en ligne des mots et en valeur des fréquences d'usage[17]. Il conçut l'analyse factorielle de correspondances qui est une méthode permettant de hiérarchiser des dépendances entre les lignes et les colonnes du tableau.

Postérieurement, des doctorants et spécialistes du domaine ont contribué à étoffer la connaissance du sujet et sa littérature via divers travaux de recherche. Des revues spécialisées en sciences humaines et sociales se sont emparées du sujet. Des initiatives ont aidé à la création de plusieurs logiciels de textométrie. Des opportunités professionnelles furent développées ; la possibilité de faire du quantitatif sur des données qualitatives a intéressé les secteurs des études et du marketing. Plusieurs laboratoires de recherche furent créés dans les universités françaises.

---

[16] Ça prend combien de temps de lire un livre ? La durée moyenne de lecture des grands classiques passée au crible | Clémence Jost
Archimag.com | publié en septembre 2014 | consulté en octobre 2018

[17] Retour aux origines de la statistique textuelle : Benzécri et l'école française d'analyse de données Valérie Beaudouin
Archives-ouvertes.fr | publié en Octobre 2016 | consulté en octobre 2018

Plus tard, des travaux spécifiques tels que ceux des Professeurs André Salem et Max Reinert ont fait valoir de nouvelles perspectives sur lesquelles une partie de la mission décrite ici est basée. Ces travaux ont contribué à la création des logiciels Alceste et Lexico dont une des spécificités est l'identification des segments répétés qui sont des séquences de mots revenant plusieurs fois dans un corpus.

Depuis 1991, les journées internationales d'analyses textuelles réunissent tous les deux ans une communauté de chercheurs essentiellement français, mais aussi italiens et espagnols, enrichissant le corpus de la revue Lexicométra.

II.5.3.3. L'offre éditeur française

Réputée dans les outils de Traitement Automatique du Langage Naturel, la France dispose d'un savoir-faire en la matière représenté par un marché d'éditeurs à la fois historiques et récents.

Initiés par la Défense Américaine pendant la période de la guerre froide durant les années 50 afin de réaliser des traductions automatiques, la discipline s'est développée plus tard grâce aux travaux d'Alan Turing. L'augmentation de la puissance informatique et l'enrichissement des ontologies ont contribué à la création d'outils plus performant puis à un développement du marché du TAL[18].

L'expertise française représentée depuis la fin des années 50, par l'Association pour le Traitement Automatique des Langues, publie sur sa revue des résultats de travaux de recherches internationaux et anime des conférences et journées d'études.

Le marché français porté par la dynamique de recherche fut initialement composé par de grandes entreprises informatiques et industrielles. Au courant des années 90, Bull, Dassault, Aérospatiale ou EDF partageaient avec les américains IBM et Xerox le marché de la traduction automatique.

Plus tard, avec développement de l'Internet, les éditeurs d'Antidot, de Sinequa ou de TEMIS commencèrent à occuper le terrain de l'offre de moteur de recherche. Ils proposèrent dans un second temps les outils de fouille de données et d'extraction d'information.

Le développement récent du stockage et de la puissance de calcul dans le nuage a offert la possibilité aux éditeurs d'évoluer du NLP (Natural Language Processing) au NLU (Natural Langage Understanding)[19] contribuant à l'essor de nouveaux produits permettant de catégoriser automatiquement des textes et le dialogue homme-machine.

L'offre en 2017-2018 est proposée par des éditeurs historiques et de nouveaux entrants (Proxem, Syllabs, Synomia, Yseop)[20].

[18] Traitement automatique du langage naturel
fr.wikipedia.org | consulté en novembre 2018

[19] NLP, NLU, NLG and how Chatbots work | Anush Fernandes
chatbotslife.com | publié en novembre 2017 | consulté en novembre 2018

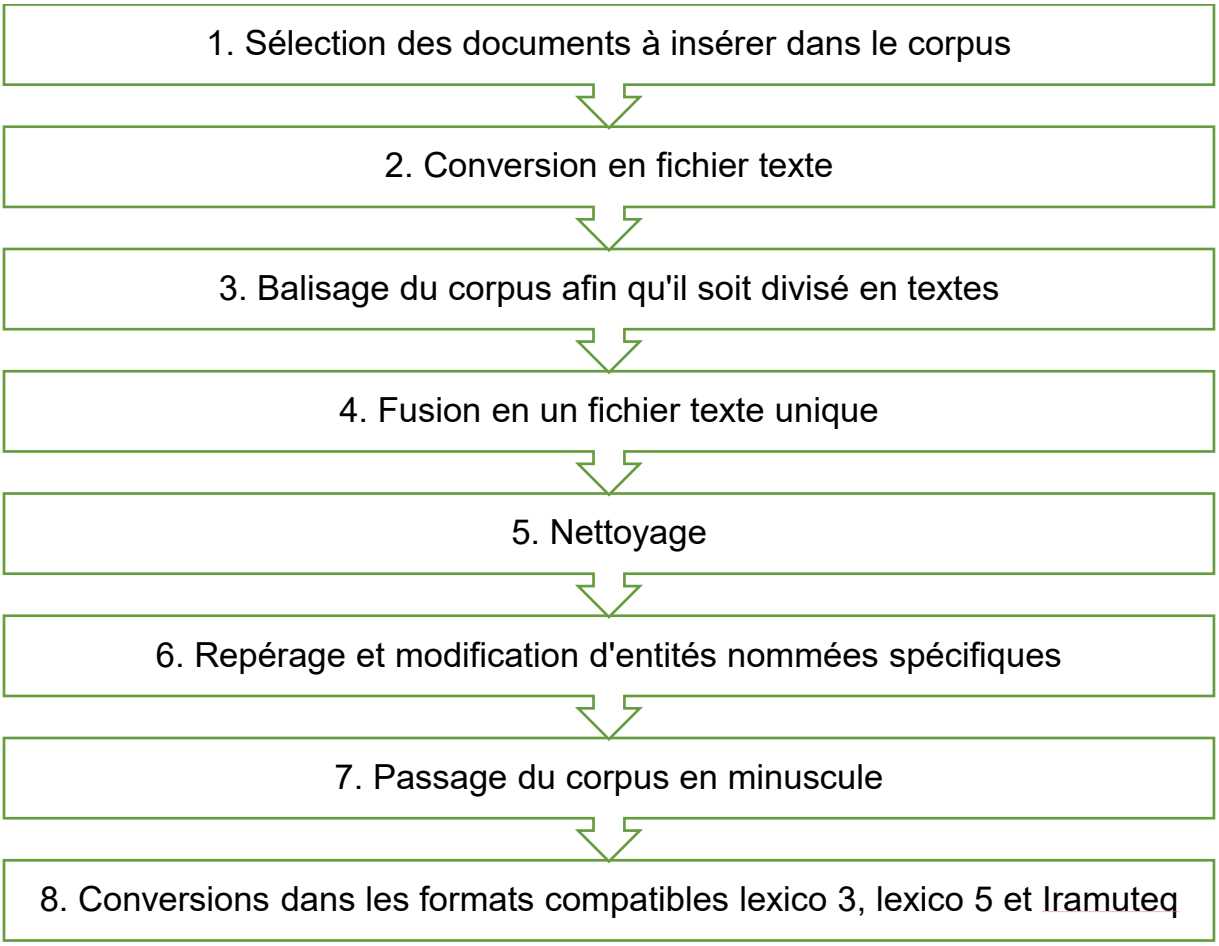
[20] LE TAL, FILS PRODIGUE DE L'IA | Gaëlle Recourcé
forum.gfii.fr | publié en décembre 2017 | consulté en novembre 2018

II.5.3.4. Préparation du corpus

Un corpus de document n'est généralement pas directement exploitable dans les outils de text-mining. Il est nécessaire de le préparer afin de le rendre compatible avec les outils.

Un tel volume de de documents implique l'automatisation de la conversion dans les formats compatibles avec les logiciels de text-mining. En s'inspirant de nombreuses ressources sur les forums d'informaticiens Stackoverflow et Microsoft, nous avons pu créer des scripts capables de préparer un corpus de plusieurs milliers de fichiers en quelques heures.

Ci-dessous sont décrites les étapes que les scripts réalisent :



n. Etapes techniques de préparation du corpus

Commentaires :

Etape 1

Etape préalable à la constitution du corpus, il s'agit de choisir les types de fichiers convertibles. Majoritairement constitué de fichiers PDF, seuls ces fichiers ont été sélectionnés à la préparation du corpus.

Etape 2

Via un outil, chaque fichier PDF est converti en format texte.

Etape 3

Au début de chaque texte, des balises permettant d'identifier le texte par son type, son année de publication et son nom sont ajoutés

Etape 4

Via plusieurs opérations de fusion, cette étape consiste en la fusion de tous les fichiers textes en un seul.

Etape 5

Le nettoyage consiste en la suppression de divers caractères spéciaux pouvant mettre en erreur les programmes de text-mining.

Etape 6

Plusieurs mots fréquemment utilisés peuvent orienter les résultats statistiques dans une direction inadaptée. Ainsi l'entité nommée fréquemment utilisée « Crédit Agricole » composée de deux mots a un sens différent des mots « crédit » et « agricole » parfois utilisés seuls. Cette étape permet de répérer ces compositions de mots afin de les fusionner et de les repérer facilement. « Crédit Agricole » devient grâce au traitement « créditagricole », excluant ainsi une compréhension inadaptée de l'usage de ceux deux mots.

Etape 7

Etape transitoire et facultative en fonction du corpus, elle permet de transformer ses caractères en minuscule afin de ne pas différencier les mots en fonction de leur casse. Cette étape a du sens dans le présent contexte, mais n'est pas forcément conseillée dans un contexte différent. Elle n'est cependant pas toujours conseillée, car l'usage des variations de la casse peut permettre la compréhension d'un sens particulier[21].

Etape 8

La conversion au format approprié consiste en :

- La personnalisation du balisage des textes en fonction du logiciel de text-mining,
- La conversion du fichier dans un format de lecture adapté au logiciel : ANSI ou UTF8

Cette étape permet de rendre le corpus exploitable par les outils de text-mining.

[21] Discours d'entreprise et organisation de l'information Apports de la textométrie dans la construction de référentiels terminologiques adaptables au contexte | Frédéric Erlos
Thèse présentée et soutenue en Novembre 2008 | Page 444
archives-ouvertes.fr | Consulté en décembre 2018

II.5.3.5. Outils et techniques informatiques avancés

Pour des raisons pédagogiques et financières, nous avons exploité les outils développés par les laboratoires de recherche en text-mining. Il existe de nombreuses initiatives et nous avons choisi de travailler avec Lexico et Iramuteq.

Lexico

Logiciel de statistiques textuelles initialement conçu à l'ENS Fontenay-Saint-Cloud par l'équipe de lexicométrie et textes politiques dirigée par Michel Tournier et André Salem, le projet a continué au sein du SYLED-CLA²T (*Système Linguistiques Énonciation Discursivité - Centre d'Analyse Automatique des Textes)* de l'Université Sorbonne Nouvelle - Paris 3.

Ce programme, qui repose sur le seul comptage des formes graphiques brutes (non lemmatisées) a pour spécificité le fractionnement du corpus en partitions à des fins de comparaison de ses diverses parties. Il permet d'extraire les fréquences d'usage des mots et groupes de mots via la génération de tableaux croisés. Le langage de la lexicométrie nomme ces analyses « Tableaux Lexicaux Entiers » regroupant les fréquences d'usage des « formes » et « Tableaux des Segments Répétés » regroupant les fréquences d'usage des « segments ».
Lexico donne aussi un indice permettant de savoir la surutilisation ou la sous-utilisation d'un terme par partition du corpus via l'analyse des spécificités. [22]

Iramuteq

Projet initié courant 2009, Iramuteq est une interface graphique de « R » développé au sein du laboratoire Lerass. Iramuteq communique des demandes d'analyses statistiques textuelles à R qui utilise des packages dédiés aux différents types de statistiques. « R » renvoi les données traitées et des schémas à Iramuteq. [23]

Iramuteq génère des données et graphiques dont les classifications hiérarchiques descendantes basées sur la méthode Alceste de Max Reinert.

En se basant sur un lexique, Iramuteq est caractérisé par sa capacité à lemmatiser un corpus et différencier les formes actives des formes supplémentaires.

---

[22] Lexico 3. Outil de statistique textuelle - manuel d'utilisation
Cédric Lamalle, William Martinez, Serge Fleury, André Salem, Béatrice Fracchiolla, Andrea Kuncova, Aude Maisondieu
lexi-co.com | publié en février 2003 | consulté en novembre 2018

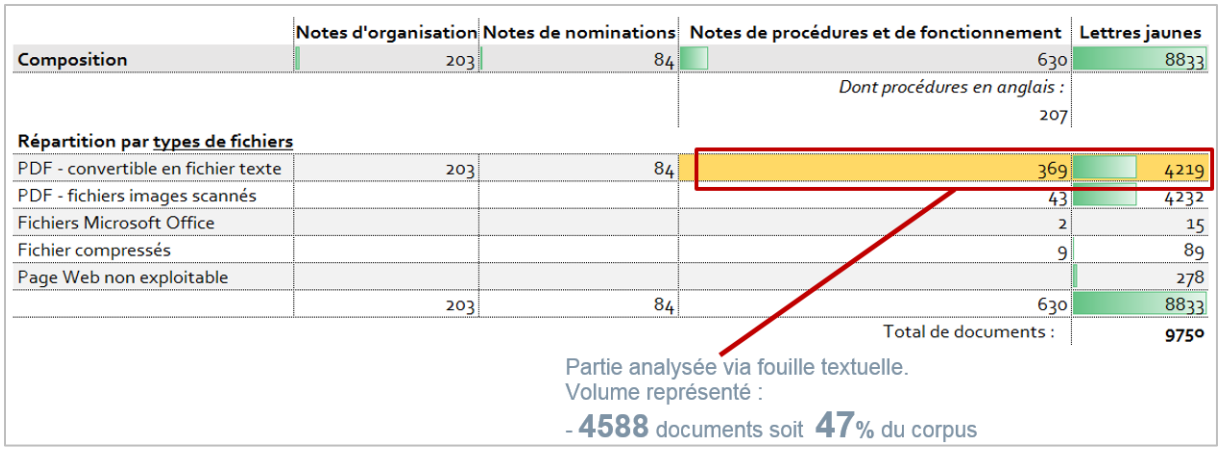
[23] Comment préparer l'analyse de textes de sites Web grâce à la lexicométrie et au logiciel Iramuteq ?
Daniel Pélissier | Présence numérique des organisations
presnumorg.hypotheses.org/187 | publié en avril 2016 et mis à jour en mars 2017 | consulté en novembre 2018

Données entrantes

Le corpus présente diverses spécificités en terme de type de fichiers. Il est constitué des documents émis par les Affaires Générales de 2002 à nos jours.

Il comporte :

- Des fichiers PDF dont il est possible de convertir le contenu en format texte,
- Des fichiers PDF comportant des documents scannés n'ayant pas bénéficié de traitement de reconnaissance des caractères. Ces documents sont en proportion majoritaire de 2002 à 2005. Puis ils reculent nettement pour ne plus être présents sur les dernières années.
- Des fichiers Microsoft Office,
- Des fichiers compressés,
- Et des pages web confidentielles non exploitables via la lexicométrie.



	Notes d'organisation	Notes de nominations	Notes de procédures et de fonctionnement	Lettres jaunes
Composition	203	84	630	8833
			Dont procédures en anglais : 207	
Répartition par types de fichiers				
PDF - convertible en fichier texte	203	84	369	4219
PDF - fichiers images scannés			43	4232
Fichiers Microsoft Office			2	15
Fichier compressés			9	89
Page Web non exploitable				278
	203	84	630	8833
			Total de documents :	**9750**

o. Corpus : répartition des types de documents selon leur format

La fouille textuelle a été réalisée sur 47% du corpus composé des fichiers PDF convertibles au format texte.

Données sortantes

Iramuteq et Lexico produisent en sortie plusieurs données d'analyses brutes et des graphiques. Nous avons manipulé les données des « tableaux lexicaux entiers » et des « tableaux des segments répétés » afin d'en ériger un sens.

La compréhension du corpus passe par l'interprétation de ces analyses directement produites par les logiciels ou par transformation des données. Ainsi, nous avons interprété les analyses factorielles de correspondances, les classifications hiérarchiques descendantes et avons créés une analyse statistique montrant l'évolution d'usage des mots les plus utilisés dans le corpus.

Usage des effectifs relatifs

Les effectifs (ou fréquences) relatifs permettent de comparer des répartitions de population de tailles différentes. En appliquant des fréquences relatives d'usage de mots ou de syntagmes sur un corpus, il est facile de comparer leur représentativité par texte.

Exemple :

Le syntagme « appétence au risque » est mieux représenté s'il apparait trois fois dans un document de deux pages que s'il apparait cinq fois dans un document de cent pages. En divisant son nombre d'apparitions par le nombre total de mots, on obtient une fréquence relative qui, multipliée par 1000 donne une évaluation en ‰. Dans le document de

deux pages, il apparaitra à une fréquence relative de 30 ‰ tandis que dans le document de 100 pages, il apparaitra à une fréquence relative de 2‰.

L'usage des effectifs relatifs est un mode de calcul sous-jacent ayant permis plusieurs analyses qui seront détaillées sur la suite du mémoire.

Sous-corpus par tranches d'années

L'ensemble du corpus englobe des documents issus de périodes marquées par divers événements. Assujettis à diverses crises économiques les législateurs durent imposer de nouveaux cadres règlementaires. Cette diversité d'actualité marque d'autant plus les documents de chaque période. Incluse dans les discours présents dans les documents publiés par les Affaires Générales entre 2002 et 2018, cette diversité tend à faire disparaitre les spécificités dans la masse documentaire. Cet aspect a été visible dans les analyses portant sur l'intégralité du corpus.

C'est pourquoi, certaines analyses ont fait l'objet de segmentation dans le temps découpé par tranches de trois ou quatre ans.

Analyse Factorielle des Correspondances

L'Analyse factorielle des correspondances (AFC) est une méthode développée par Jean-Paul Benzécri permettant de hiérarchiser des relations statistiques pouvant exister entre individus placés en ligne et des variables en colonnes dans un tableau. Des présentations graphiques de ces calculs ont été créé facilitant leur compréhension.

Méthode Reinert

La méthode Reinert permet de regrouper les termes utilisés en thématiques et comprendre leurs dépendances et relations. Elle permet ainsi de regrouper les mots proches afin d'en comprendre le sens.

Illustrations – AFC et Méthodes Reinert

Au regard de l'intérêt des utilisateurs pour les documents récents (voir l'illustration g), nous avons réalisé des illustrations AFC et CHD sur la période 2015 - 2018.

Il en est ressorti plusieurs points marquants :

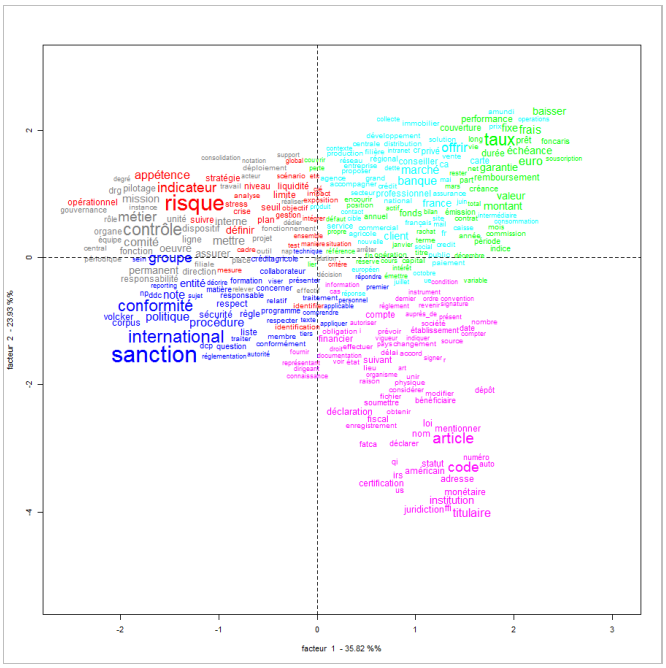
Les thématiques d'offres sur la période sont statistiquement au nombre de six. Elles sont représentées par diverses couleurs et les deux illustrations démontrent leurs multiples adhérences.

En résumé :

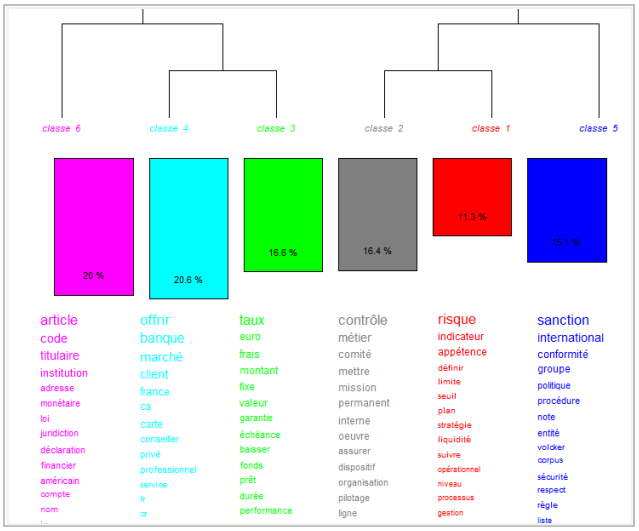
Les problématiques juridiques impliquent des mises à jour importantes du code monétaire et financier. Ce qui impacte les problématiques de liquidité et la nouvelle offre bancaire.

Le respect des sanctions internationales passe par diverses opérations de contrôle afin de quantifier les risques et d'assurer la conformité de l'activité bancaire.

Cette première vision offre une compréhension globale des enjeux du groupe bancaire par l'interprétation des statistiques érigées sur la version la plus récente du corpus.



p. Illustration : AFC



q. Illustration : CHD

La présente interprétation montre que les sujets émanant du corpus sont liés à l'actualité juridique et économique de l'entreprise. Nous y distinguons des problématiques ayant de multiples adhérences.

Diachronie : première apparition des dix formes les plus utilisées par texte

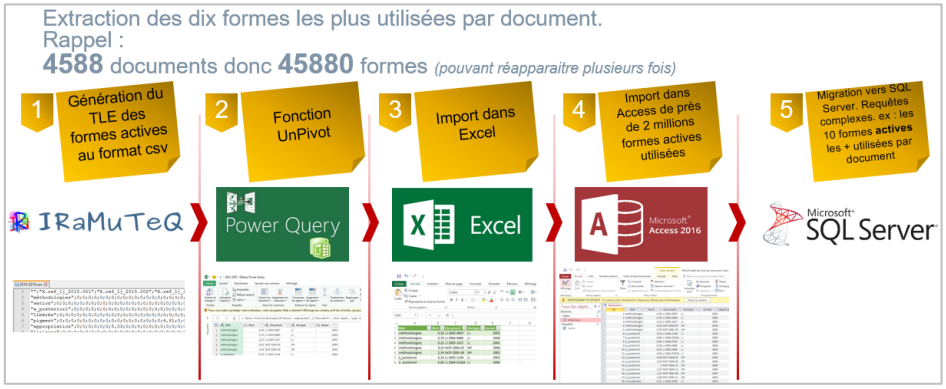
Plusieurs études ont mis en évidence les variations de langage dans le temps. Le « temps lexical » démontre comment le vocabulaire peut diachroniquement évoluer faisant apparaitre des spécificités sur des périodes déterminées. Via l'analyse des spécificités, lexico 3 permet de déduire la sous-utilisation ou la surutilisation d'un segment ou d'une forme sur une période représentée par différents textes du corpus.[24]

Afin de vérifier l'hypothèse envisageant que l'usage des mots d'un champ lexical d'une thématique évolue diachroniquement, il est intéressant de savoir à partir de quel moment ces mots sont fréquemment utilisés.

Suivant cette théorie et au regard du volume documentaire, nous avons réalisé un ensemble de requêtes permettant de savoir quand un mot apparait pour la première fois au top dix des mots les plus utilisés par texte.

Pour ce faire nous avons conçu et suivi le processus décrit ci-après depuis le TLE généré par Iramuteq.

En lexicométrie un mot est symbolisé par une chaine de caractères différents du caractère espace et des autres délimiteurs de formes comme les signes de ponctuation. Cette chaine de caractères est appelé une « forme ».



r. Processus d'extraction des dix des formes les plus utilisées par document

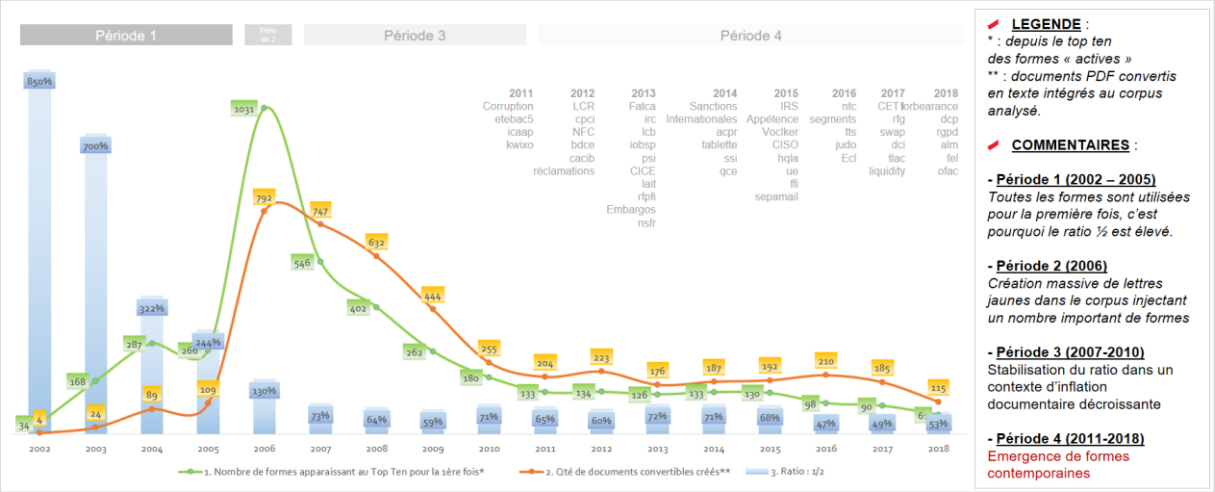
Etape 1	Après avoir intégré le corpus dans le logiciel Iramuteq, une analyse factorielle des correspondances génèrant un TLE représentant les fréquences relatives des formes par document est réalisée.
Etape 2	Via Power Query, outil d'extraction de données intégré à Microsoft Excel, le TLE est « dépivoté » afin de mettre en colonne trois données : la forme, la fréquence et le texte, représentant la fréquence relative d'utilisation d'une forme par texte
Etape 3	Etape intermédiaire de mise en conformité des données, l'import dans Excel permet de préparer la donnée brute issue de Power Query
Etape 4	Importation dans le SGBD « Access »
Etape 5	Importation dans le SGBD « SQL Server » afin de réaliser la requête complexe

3. Etapes d'extraction du top ten des formes les plus utilisées par document

[24] Mettre en évidence le temps lexical dans un corpus de grandes dimensions : l'exemple du Parlement européen | Sascha Diwersy, Giancarlo Luxardo | JADT 2016 archives-ouvertes.fr | publié en septembre 2016 | consulté en novembre 2018

Il est nécessaire d'importer les données dans le SGBD serveur « SQL Server » car son moteur plus puissant permet de réaliser la requête extrayant les dix formes les plus utilisées par texte en quelques minutes, ou Access a besoin de plusieurs dizaines d'heures.

Via la requête réalisée, nous avons construit de nouveaux indicateurs permettant de comprendre comment le langage a évolué entre 2002 et 2018.



s. Evolution diachronique des dix formes les plus utilisées

3 indicateurs annuels ont été conçus :

- Nombre de formes apparaissant pour la première fois au top dix des formes les plus utilisées
- Nombre de documents annuellement générés
- Ratio divisant le premier indicateur sur le second

En commentaires sont identifiées certaines formes démontrant la pertinence de l'analyse.

Commentaires :

Si les deux premières périodes graphiquement identifiées démontrent une injection massive de formes dans le top dix, les suivantes sont l'opportunité de mieux comprendre les nouvelles problématiques par l'émergence d'utilisation de nouveaux termes métiers.

Exemples :

Nous pouvons constater qu'à partir de 2014, les formes « sanctions » et « internationales » sont très utilisées dans le contexte où BNP Paribas est condamné à s'acquitter d'une amende de plusieurs milliards de dollars suite au non-respect des lois américaines sur l'usage du Dollar avec ses pays sous embargos. Nous en déduisons une intensification de la connaissance du sujet « sanctions internationales » par les autres acteurs du marché depuis cet événement.

En 2018 la forme RGPD apparait. Ce terme est l'acronyme de Règlement Européen sur la Protection des Données. Devenu applicable en 2018, les documents y font fortement référence depuis son application.

Ces indicateurs démontrent que les usages des termes sont en rapport avec leur époque. Ils révèlent par ailleurs, que dans le cadre d'une mise à disposition de nouveaux moyens d'accès à l'information, il est nécessaire de régulièrement s'adapter à l'actualité économique, juridique et organisationnelle.

II.5.3.6. Conclusion intermédiaire : l'analyse du corpus

Face à un corpus métier structuré et organisé en types de documents, les outils de text-mining permettent d'identifier les grandes lignes de son contenu.

Il aurait été possible de faire appel à des éditeurs de logiciel, mais afin de comprendre le fonctionnement des mécanismes internes au text-mining et pour des raisons pédagogiques, il a été fait usage d'outils issus des travaux universitaires.

En faisant usage de ces outils il est possible d'utiliser les concepts initialement développés en laboratoire dans un milieu professionnel. De nombreuses étapes techniques de mise en conformité du corpus restent à réaliser et nécessitent un bon niveau de connaissance en informatique. Effectivement, il fut nécessaire de concevoir ou personnaliser des scripts et d'utiliser des fonctions avancées en bases de données et en SQL.

La majeure partie du corpus des Affaires Générales ne peut être analysée, vue qu'elle n'est pas exportable dans les outils de text-mining. Cependant, la partie analysée est la plus récente et aussi la plus consultée.

Le corpus est marqué par différentes époques. Suite aux diverses crises financières, le législateur a contribué à une inflation règlementaire se traduisant par de nombreux textes déclinant leurs applications dans l'entreprise.

Des découpages du corpus par périodes ont mis en évidence les évolutions, la diversification et les adhérences des sujets.

L'usage du text-mining a révélé des contingences proximités entre les sujets. Certains documents se ressemblent et peuvent contenir divers sujets abordés dans les thématiques d'offre du corpus ou de demande des utilisateurs.

Si dans la première partie de l'analyse de l'audience il a été constaté que les demandes des intranautes portent sur plusieurs thématiques contemporaines identifiées, il est important de souligner que l'offre documentaire récente s'en rapproche.

III. Préconisations & Spécifications

III.1. Enjeux contemporains - anticiper le besoin d'accès à une information toujours plus complexe

III.1.1. Les modes d'accès classiques aux documents

En répondant au foisonnement bibliographique, documentaire et informationnel, les différents langages documentaires proposent de rassembler les sources de connaissances partageant plusieurs points communs, afin d'en faciliter l'accès et optimiser la recherche.

Les langages documentaires se répartissent entre classification et description. Tandis que la classification permet une hiérarchisation des thématiques, la description offre la possibilité d'associer un document de toute nature à un groupe de termes ou de mots clés.[25]

Alors que les bibliothèques utilisent diverses méthodes de de classement telles Dewey ou Rameau pour la BnF, les thésaurus permettent de catégoriser dans un langage évolutif, structuré et combinable diverses médias.

Evolution du classement numérique, les ontologies modélisent un domaine de connaissance par l'usages de relations de sens entre objets du domaine concerné. Les ontologies sont employées divers domaines tels que l'intelligence artificielle, le web sémantique, l'informatique biomédicale et l'architecture de l'information.

Les présents portails offrent des modes d'accès classiques basés sur ces langages documentaires, maitrisés par le documentaliste contemporains mais sous-utilisés par les utilisateurs.

III.1.2. De l'utilisateur au consommateur

Le développement du numérique et l'inflation documentaire a entrainé des mouvements de transitions des interfaces d'accès l'information.

Par mimétisme sur les usages personnels, les utilisateurs ciblent des thèmes, consomment les médias dont ils ont besoin en vue d'en découvrir davantage sur un sujet ou pour orienter leurs décisions.

En réponse à cette demande, nous pouvons citer l'exemple de Canal +, qui a créé des accès thématiques sur son application My Canal déclinées sur smartphone, tablette et téléviseurs connectés. Molotov TV classe les émissions de catch-up TV par thèmes plutôt que par chaines.[26]

Si le moteur de recherche est toujours présent sur les portails, les accès thématiques apportent une nouvelle réponse au consommateur de médias. Ils répondent à plusieurs besoins : visualisation et choix de d'informations puis sérendipité.

Les accès thématiques doivent cependant répondre à un besoin plutôt qu'à un impératif de classement. Le documentaliste contemporain doit produire des thèmes en rapport avec l'actualité et y classer des médias pertinents.

---

[25] Les langages documentaires, principes, histoire et perspectives
Support de cours | Publié en janvier 2018 | Loïc Lebigre

[26] ORLM-255 : Canal+, Netflix, YouTube, Apple, demain la TV !
onrefaitlemac.com | Publié en mars 2017 | Consulté en Décembre 2018

III.1.3. Adhérences & diversification des sujets

Rassemblant près de 10000 documents sur 16 ans, le corpus des Affaires Générales relate en partie l'histoire du Groupe via l'actualité de son environnement. Il comporte un nombre important de documents sur les huit premières années qui, pour des raisons de rationalisation rassemble moins de publications par la suite. L'analyse par groupe d'années du corpus via le text-mining a révélé une croissance de la quantité de sujets couverts. Si les documents étaient nombreux et traitaient de peu de sujets au départ, ils sont récemment bien plus imposants et peuvent rassembler de nombreux sujets.

Cet état de fait marque l'époque du décloisonnement des entreprises se traduisant dans les documents Groupe publiés sur l'intranet. Les groupes de travail rassemblant de nombreux experts agrègent plusieurs sujets contingents faisant consensus dans un document. Cette complexité nouvelle peut rendre difficile l'accès aux informations à la fois dispersées dans plusieurs documents, chacun traitant de divers sujets.

III.1.4. Anticiper le besoin des travailleurs de la connaissance

Les travaux d'analyse de l'audience et du corpus ont mis en évidence une relation entre le temps et les requêtes d'accès aux connaissances.

Les utilisateurs sont généralement en demande d'informations récentes provenant de l'actualité interne et externe à l'entreprise. Les recherches sont marquées par l'actualité juridique, économique et l'offre de produits émanant de l'environnement du groupe. Evoluant au gré de l'actualité, les documents comportant les informations recherchées deviennent moins nombreux, mais comportent une diversité croissante de sujets.

Dans l'article « *Facettes et systèmes d'information. Une approche focalisée sur un besoin de savoir pour agir* », Francis Beau exprime que la fonction d'indexation s'assimile mieux à une logique de vendeur répondant à un besoin client qu'à un besoin d'organiser un stock de documents parfaitement répertorié. L'indexation consiste en l'identification de *« différents thèmes utiles à l'exploitation qu'un document permet de capitaliser dans une mémoire »*.[27]

Similairement, le besoin dégagé par les travaux réalisés précise qu'il est en constante évolution puis de nature spécifique et limité dans le temps. L'évolution des sujets et de la terminologie nécessite une adaptation constante de l'offre d'accès.

Partant de ces postulats, le portefeuille de projets propose la conception de nouveaux accès à des thématiques de connaissances en mouvement.

[27] Facettes et systèmes d'information.
Une approche de la classification focalisée sur un besoin de savoir pour agir | Francis Beau
Lavoisier | Les cahiers du numérique | 2017/1 Vol. 13 | pages 115 à 142
Cairn.info | publié en 2017 | consulté en novembre 2018

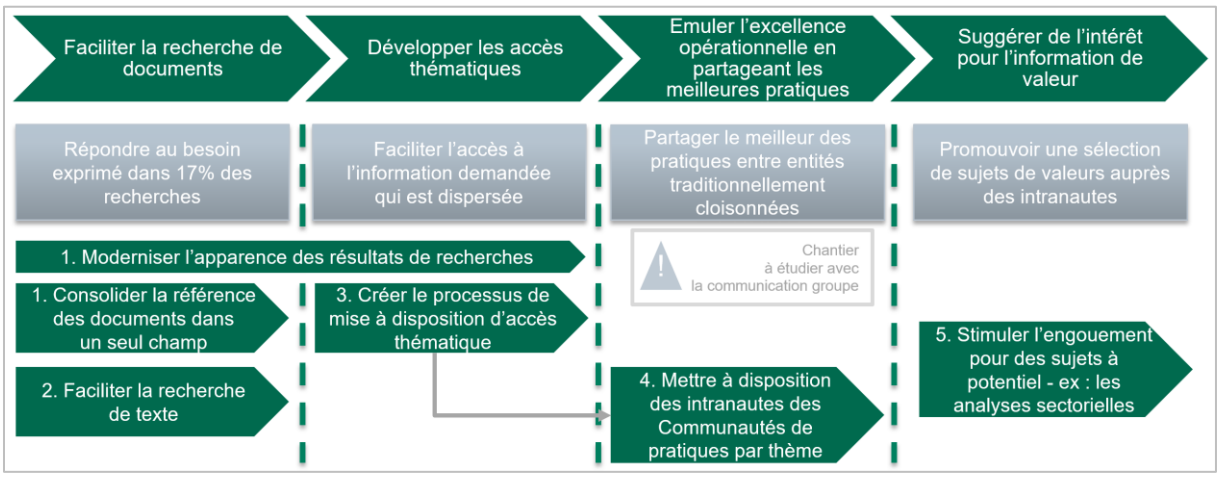
III.1. Aligner l'offre sur la demande

Pour donner suite aux démarches d'analyse de l'audience et du corpus, un portefeuille de projets *adaptable* comprenant diverses améliorations a été proposé. Il offre de nouveaux moyens d'accès et de partage des informations recherchées par les utilisateurs. Il est qualifié d'adaptable dans la mesure ou le commanditaire peut choisir si un projet doit être mené.

Sachant que les intranautes recherchent (voir illustration n) :

- Des documents d'après leur référence,
- Des informations contenues dans les documents en utilisant des termes métiers regroupés en thématiques de recherche,
- Des connaissances associant des informations et des savoir-faire,

le portefeuille a été structuré en quatre projets comprenant plusieurs actions :



t. Proposition de portefeuille de projets

Les quatre projets proposés sont basés sur le présent intranet des Affaires Générales. Modernisé par « touches successives » il répond au besoin des utilisateurs.

Le schéma précise :

- Dans sa partie supérieure : les projets proposés sous forme de flèche d'action,
- Dans les cases grises, en quoi les projets répondent aux besoins
- Et dans la partie inférieure, le plan d'actions ou de sous projets nécessaires à la transformation.

Les projets et leurs adhérences sont décrits dans la suite de la proposition.

III.2. Contenu du portefeuille de projets

III.2.1. Optimiser la recherche de documents

Répondant à un besoin identifié de retrouver les documents, ce projet consiste en la modernisation de l'interface graphique de l'intranet et la valorisation de la métadonnée « référence du document ».

III.2.1.1. Genèse du besoin

L'analyse de l'audience a mis en évidence que les utilisateurs recherchent sur un an des documents par leur référence dans 17% des cas. C'est un besoin statistiquement identifié et qui a été manifesté au cours de diverses réunions de travail.

III.2.1.2. Analyse d'opportunités

Souhaitant répondre à la demande, la DSI a proposé une modernisation de la recherche avancée. Cependant, les résultats de l'analyse de l'audience n'ont montré que peu d'intérêt pour la recherche avancée ne laissant que quelques centaines de chargements sur un an alors qu'environ 30000 recherches ont été comptabilisées.

III.2.1.3. Propositions

Trois actions ont été proposées :

Valorisation de la métadonnée correspondante à la référence du document

Les différentes bibliothèques de documents SharePoint composant le corpus, disposent d'une métadonnée « nom du document » normalisée. Les recherches de documents étant basées sur cette information, sa qualité doit être préalablement consolidée par des travaux de croisement des données issues d'un référentiel et des données de bibliothèques SharePoint.

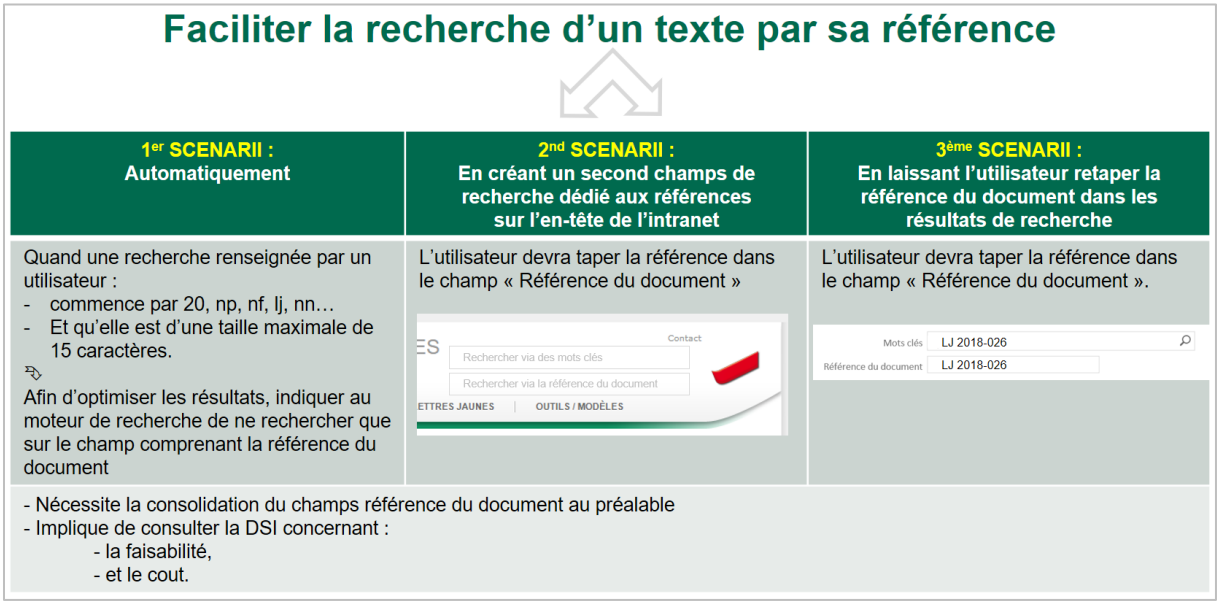
Après comparaison avec le référentiel, les données seront consolidées.

Moderniser la recherche par la proposition de plusieurs scénarii :

La recherche par l'utilisateur est l'action consistant à renseigner l'information demandée dans le moteur de recherche.

Sur l'intranet, les utilisateurs instruisent des mots dont la référence du document afin de trouver le document qu'ils souhaitent consulter.

Afin de répondre à la modernisation du besoin, trois scénarii classés d'après leur complexité sont proposés. Le commanditaire pourra en choisir un d'après ses contraintes opérationnelles.



u. Scénarii de modernisation de la recherche

Le premier scénario implique la programmation du moteur afin qu'il ne recherche que sur la métadonnée « nom du document », lorsque l'intranaute instruit des informations correspondantes au nom du document. Cette solution, bien qu'élégante nécessite des développements et des paramétrages complexes du moteur de recherche.

Le second scénario consiste en la superposition d'un nouveau champ de recherche via la référence du document à proximité du champ de recherche initial. Cette proposition facilite le travail de la DSI mais pourrait entrainer les utilisateurs dans la confusion.

Le troisième scénario propose d'ajouter le champ de recherche dans les résultats de recherche uniquement. Cette proposition implique que l'utilisateur instruise une nouvelle fois le nom du document dans le champ dédié si le document recherché n'est pas trouvé par l'usage préalable d'une recherche standard. Cette proposition, peu onéreuse serait fonctionnelle mais risque de ne pas répondre au besoin des utilisateurs au regard de son ergonomie sommaire.

Simplifier les résultats de recherches

La page des résultats de recherche présente des informations accompagnées de facettes permettant de filtrer les résultats suivant plusieurs métadonnées historiques. Certaines de ces métadonnées représentent des informations n'étant plus utilisées sur les documents récents.

Parallèlement, nous avons pu remarquer que les utilisateurs souhaitent retrouver les informations récentes et comprendre si les notes de procédures et de fonctionnement sont en vigueur ou abrogées. En ajoutant la possibilité de filtrer la validité via une facette et en y donnant une visibilité via un jeux d'icônes significatifs, la consultation de l'intranet par les utilisateurs est facilitée. Tel le site internet publiant la législation de l'UE Eur-Lex dans sa présente version, un texte en vigueur disposera d'un feu un vert, là où un texte abrogé aura un feu rouge.

III.2.2. De nouveaux accès thématiques basés sur un processus

III.2.2.1. Genèse du besoin

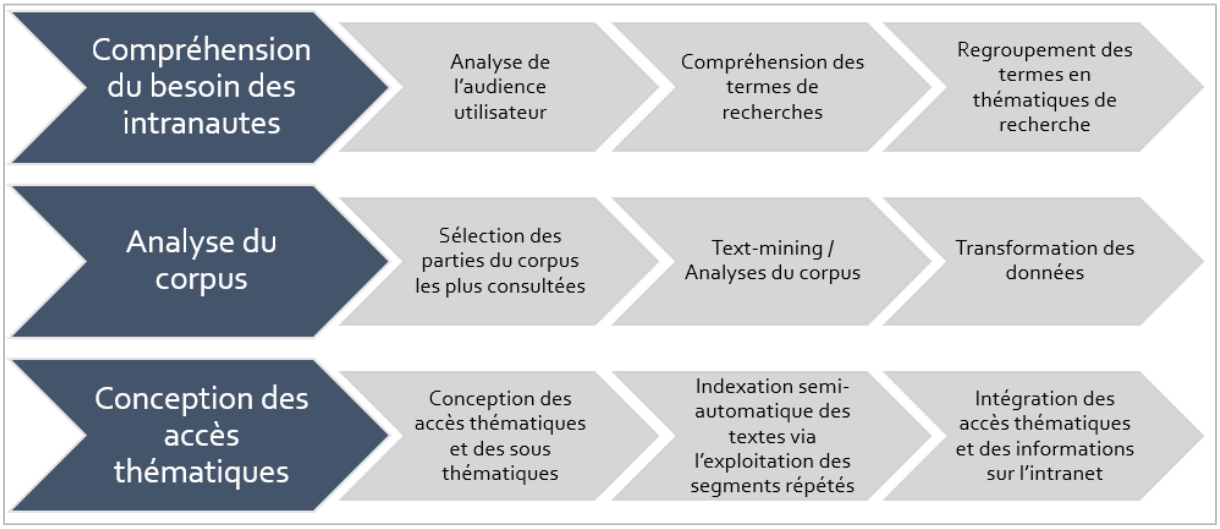
L'analyse de l'audience a mis en évidence que les termes employés par les intranautes dans le moteur de recherche sont *regroupables* en thématiques de recherches. En outre, elle a mis en perspective le fait que la demande évolue au fil de l'actualité des sujets couverts par le corpus. Un parallélisme a été constaté entre les termes fréquemment utilisés dans le moteur de recherche et les syntagmes fréquemment employés dans la partie récente du corpus.

En déduction : les utilisateurs recherchent des informations récentes qui existent dans les documents contemporains via un vocabulaire commun au corpus et aux recherches. D'après l'analyse de l'audience, les utilisateurs ne trouvent cependant pas aisément les documents en rapport avec les sujets recherchés. Les documents étant parfois transverses, ils traitent de plusieurs thématiques.

Sachant que le besoin évolue au gré de l'actualité et que les Affaires Générales sont en possession de données textuelles et d'audience provenant de diverses sources, il est opportun d'initier une démarche de conception de thématiques d'offres, via un processus à exercer régulièrement. Une fréquence semestrielle permet d'être en accord avec le besoin utilisateur en constante évolution.

III.2.2.2. Processus – de la compréhension du besoin à la conception d'une nouvelle offre

Une structuration des actions en processus global clarifie la succession de tâches à réaliser.



v. Accès thématiques : processus alignant une nouvelle offre sur le besoin des intranautes

Regroupées en trois lignes d'actions, le processus permet de comprendre le besoin afin d'en ériger de nouveaux accès thématiques :

La **compréhension du besoin des intranautes** se concentre sur l'analyse des données issues de la plateforme d'audience. Analysant la fréquentation et la terminologie employées dans le moteur de recherche, elles permettent d'identifier ce qui est recherché dans le temps et les sujets. La compréhension du besoin des intranautes se base sur l'analyse de l'audience décrite en chapitre II.5.2 de ce présent mémoire.

L'analyse du corpus implique d'utiliser les outils de text-mining afin de rendre exploitables des données provenant des textes en vue de concevoir les accès thématiques.

L'analyse du corpus est techniquement basée sur les travaux présentés dans le chapitre II.5.3 de ce présent mémoire.

Cette étape implique une sélection de textes en rapport avec l'analyse de l'audience.

La **conception des accès thématiques** implique les deux précédentes étapes en liant la demande à l'offre via la conception de nouveaux accès thématiques.

Dans le présent cas, il a été constaté que les utilisateurs ont un intérêt pour les documents 3,5 dernières années et pour les notes de procédures et de fonctionnement en vigueur.

En conséquence, le corpus extrait en vue d'une conception des accès thématique comprend :

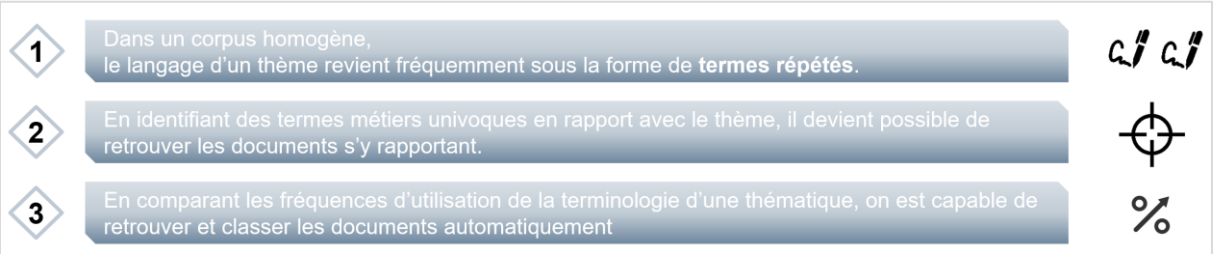
- Les lettres jaunes de 2015 à 2018
- L'ensemble des notes de procédures et de fonctionnement en vigueur.

III.2.2.3. Focus sur une application professionnelle des segments répétés

Des segments répétés à la thématisation

Nés des travaux de recherches d'André Salem, les segments répétés sont définis comme « suite d'occurrences non séparées par un délimiteur de séquence » et présents à minima par deux reprises dans un corpus. Via la modernisation des outils de lexicométrie, les suites de mots tels que « sécurité sociale » ou « pouvoir d'achat » sont repérables dès qu'ils sont utilisés plusieurs fois dans un corpus. Des expressions locutionnelles sont aussi comptées par la démarche : « de la », « afin de ».[28]

Quand leur sens est univoque, les segments répétés deviennent des syntagmes significatifs. Sachant que le champ lexical d'une thématique comprend plusieurs syntagmes significatifs, leur repérage et quantification dans texte permet d'en déduire son ou ses thèmes de façon automatisée.



w. Théorie de l'indexation automatisée par l'usage des segments répétés

Par l'association de diverses méthodes statistiques décrites dans le mémoire, des textes ont été thématisés via l'analyse de syntagmes significatifs.

[28] André Salem, Pratique des segments répétés. Essai de statistique textuelle | Simone Bonnafous Mots. Les Langages politiques / Année 1988 / n°17 / pages 243-245
persee.fr | consulté en Octobre 2018

Processus de traitement des segments répétés

Au début de l'analyse de préparation à de nouveaux accès thématiques, des textes ont été sélectionnés d'après les analyses de l'audience. Le corpus était donc composé de toutes les notes de procédures et de fonctionnement en vigueur et des lettres jaunes des 3,5 dernières années. Ensuite, le corpus a été préparé via les outils dont les scripts évoqués dans ce mémoire en vue de les analyser dans Lexico 3.

Lexico 3 a produit en sortie le T.L.E. (Tableau Lexical Entier) et le T.S.R. (Tableau des Segments Répétés) ordonnant dans un tableau croisé, des références de textes en colonne et en ligne des mots et des locutions.

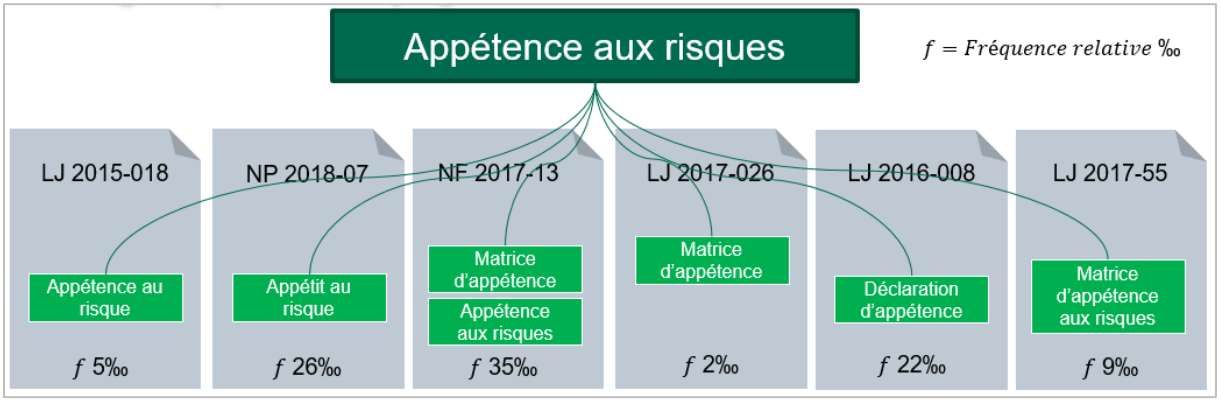
Ensuite, via Microsoft Power Query et Power BI, le tableau croisé a été désactivé sur les colonnes simplifiant les données à 3 colonnes :

- Le champ Mot / Segment : exemple « Appétence au risque »
- Le champ fréquence : exemple « 23 »
- Le champ texte : exemple « LJ 2017-93 »

Via ce triplet de données, l'information permettant de savoir la fréquence d'usage réelle d'un mot ou d'un segment répété par texte est exploitable dans une base de données relationnelle. Dans l'exemple décrit ci-dessus, le segment « Appétence au risque » est répété 23 fois dans la lettre jaune 2017-93.

Afin de pouvoir comparer l'importance de l'usage du segment entre des textes de tailles différentes, la fréquence d'usage du segment est ensuite divisée par le nombre de formes (mots) puis multiplié par 1000 exprimant ainsi son effectif relatif par texte en ‰.

Ne sont sélectionnés que les textes ayant un effectif réel d'au moins deux segments en rapport avec le thème présent dans leur contenu ; une fois n'étant pas significatif.



x. Simulation de l'apparition des syntagmes en rapport avec le sous-thème « appétence aux risques » dans divers textes du corpus

« Appétence aux risques » étant à la fois une sous-thématique et un terme métier décliné en plusieurs autres termes, nous avons reconstitué son champ lexical en vue de pouvoir compter toutes ses déclinaisons en ‰ par texte.

La simulation ci-dessus illustre comment sont repérés et comptés les syntagmes en rapport avec le champ lexical du thème « appétence aux risques ».

Ensuite, ont été additionnés les quantités d'effectifs relatifs des différents segments répétés par texte afin d'en réaliser une moyenne et classer les documents.

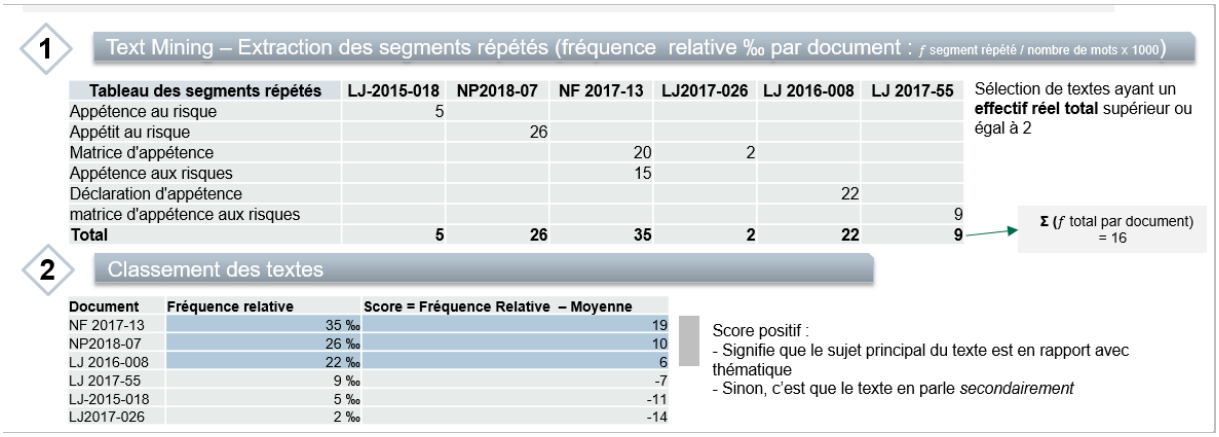


Tableau des segments répétés	LJ-2015-018	NP2018-07	NF 2017-13	LJ2017-026	LJ 2016-008	LJ 2017-55
Appétence au risque	5					
Appétit au risque		26				
Matrice d'appétence			20	2		
Appétence aux risques			15			
Déclaration d'appétence					22	
matrice d'appétence aux risques						9
Total	**5**	**26**	**35**	**2**	**22**	**9**

Document	Fréquence relative	Score = Fréquence Relative – Moyenne
NF 2017-13	35 ‰	19
NP2018-07	26 ‰	10
LJ 2016-008	22 ‰	6
LJ 2017-55	9 ‰	-7
LJ-2015-018	5 ‰	-11
LJ2017-026	2 ‰	-14

y. Du tableau des segments répétés au score de la thématique : illustration de la transformation des données

L'illustration ci-dessus précise la moyenne des effectifs relatifs des textes comprenant des segments répétés sur la thématique. Puis un calcul de score soustrayant la fréquence relative par texte à cette moyenne permet :

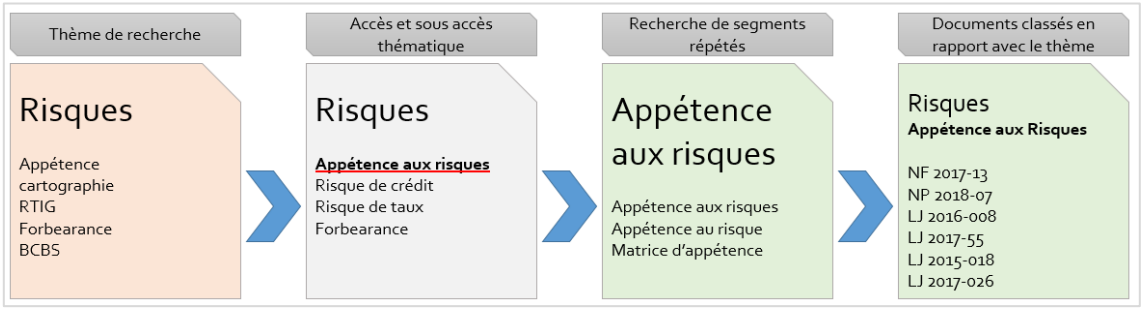
- D'identifier les documents en rapport significatif avec le thème par un score supérieur à zéro,
- Et les autres textes traitant du thème secondairement.

Réalisée via la base de données, les travaux de modélisation permettent de concevoir des thématiques avec leurs sous-thématiques, basées sur un jeu de syntagmes recherchés par texte. Le modèle de base de données et de requêtes est détaillé en annexe.

Nouveaux accès thématiques

Afin de réaliser des accès thématiques en rapport avec les demandes des utilisateurs, les termes utilisés dans le moteur de recherche ont été regroupés en thèmes de recherche.

Par la suite, les nouvelles thématiques d'accès ont été conçues depuis ces thématiques de recherches. Chaque thématique d'accès comprenant ses sous-thèmes d'accès déclinés en segments répétés.



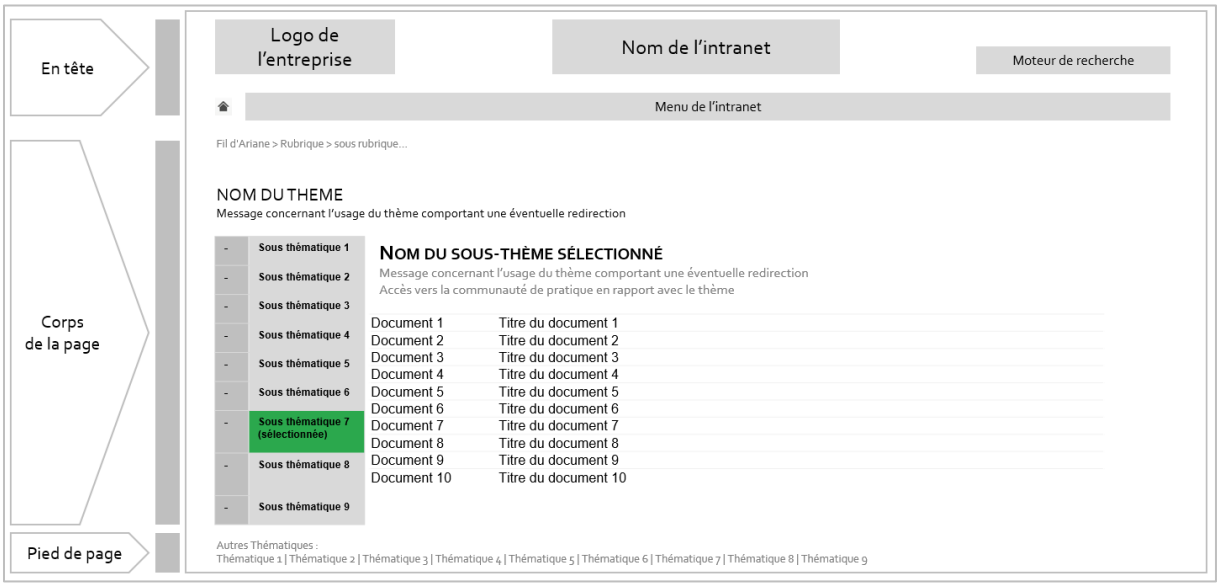
z. Audience, Thématiques, Segments et Documents : processus de classement des documents

La démarche entreprise a permis la conception de neuf nouveaux accès thématiques, déclinés en sous-thématiques d'actualité.

Nouveaux accès thématiques - Design

Etape finale de représentation de l'information, la conception de l'interface graphique des nouveaux accès thématiques implique la représentation visuelle selon divers standards de conception et d'après la charte graphique de l'entreprise.

Ont été réalisées les maquettes et les simulations de ce que seront les nouveaux accès thématiques



aa. Maquette d'une page représentant des documents d'une sous-thématique

La réalisation d'une maquette permet de visualiser les différentes parties. Ci-dessus sont décrites les trois parties composant une page de documents d'une sous-thématique.

L'usage de divers processus de représentation du contenu permet de faciliter la compréhension du projet et invite à la critique constructive et à l'adhésion au projet évitant ainsi les écueils de nombreux allers-retours chronophages.

III.2.3. Partager les savoir-faire via des communautés de pratiques

III.2.3.1. Répondre à un besoin de connaissances

L'analyse de l'audience a démontré plusieurs points dont un besoin en connaissances. L'exploration des données a révélé un usage du moteur de recherche proche des usages de recherches dans une base de connaissances. Les utilisateurs ont recherché des sujets nécessitant des retours d'expériences où le corpus ne peut apporter qu'une réponse de niveau informationnel (règlementaire, produit, nomination…). La complexité de l'ensemble des sujets associé à la dispersion des utilisateurs dans différentes filiales rend difficile l'édification des connaissances centralisées.

Des usages en gestion des connaissances ont été étudiés et partagés par le monde professionnel et universitaire. De nombreuses réussites et une littérature commence à densifier l'intérêt pour le sujet. Situé à mi-chemin entre le management de communautés et la connaissance formelle, les communautés de pratiques permettent de rassembler des usagers de la connaissance autour d'un intérêt commun érigé en thématique de discussion et d'échange.[29]

En articulant des communautés de pratiques provenant des accès thématiques, l'information provenant de ces dernières permet la confrontation de différents usages et retour d'expérience des sachants de différentes filiales du Groupe Crédit Agricole.

Les sujets demandés deviennent des thèmes, eux-mêmes déclinés en communautés de pratiques afin d'aider le Groupe à partager le meilleur de son savoir et lui permettre d'atteindre l'excellence opérationnelle.

III.2.3.2. Une conception adaptée aux contraintes règlementaires

Les contraintes de taille du Groupe Crédit Agricole impliquent des règles à décliner dans les usages des communautés de pratiques. Les organisations ont un besoin de transversalité mais il est cependant nécessaire dans certains cas de garder des frontières entre métiers ou entités.

Il est donc nécessaire d'adapter les usages des COP (Communautés de Pratiques) dans le respect de ces contraintes. Par l'usage de droits adaptés et d'une charte (précisant ces points) dont la signature autorise l'adhésion à une communauté, les contraintes y sont directement déclinées.

III.2.3.3. Un projet en gestation

Actuellement non inscrit dans la stratégie d'entreprise, ce type de projet nécessite sponsoring, commanditaires et moyens. Les idées générales ayant été transmises au commanditaire de la mission, il reste à rechercher des parties prenantes à l'adhésion du projet. La réussite de ce type de projet dépend de son orchestration. Une identification pertinente des parties prenantes permet de démarrer ce type d'approche facilement.

---

[29] Communautés de pratique et performance dans les relations de service, cas des "Front-Office". Quels enseignements pour la GRH ? | Lamine Mebarki
Thèse soutenue en 2011 | consultée en novembre 2018
archives-ouvertes.fr

III.2.4. Stimuler l'engouement pour les documents de valeurs (analyses sectorielles)

L'intranet des Affaires Générales dispose de nombreuses analyses sectorielles éditées en interne. Ces informations à valeur ajoutée permettent aux banquiers de diverses régions et métiers de répondre aux besoins de leur clientèle. Ces informations contenues dans ces documents ont une valeur certaine auprès des forces commerciales de l'entreprise.

Via divers dispositifs de consultation de ces documents, les banquiers sont mieux préparés aux réussites commerciales :

- D'une part en ajoutant un accès thématique dédié aux analyses sectorielles, leur accès facilité.
- D'autre part en choisissant une méthode marketing adaptée, ces études peuvent être suggérées en vue d'être consultées.

IV. Conclusion

IV.1. Un sujet contemporain en réponse à la complexité et à l'infobésité

La conception d'accès thématiques en rapport avec le contexte d'entreprise et le besoin des utilisateurs répond au besoin croissant de mise à disposition d'informations en mouvement.

Les cycles économiques des entreprises s'écourtent, en conséquence de quoi il y a une nécessité permanente à apporter une réponse informationnelle en rapport avec l'actualité.

Parallèlement, les documents et sujets traités deviennent complexes du fait de l'inflation règlementaire et du décloisonnement de l'organisation. Un même document pouvant traiter plusieurs sujets.

Le moteur de recherche ne peut apporter à lui seul une réponse adaptée à ce besoin d'information.

La thématisation et le partage des connaissances des sujets en accord avec leur temps offrent un nouveau moyen d'accéder et de partager l'information et les connaissances.

IV.2. Etre dans la conformité du besoin

En ayant ouvert un champ de réflexion sur la conception d'accès thématiques basés sur une analyse de l'audience et le text-mining, nous avons offert la possibilité de comprendre des besoins spécifiques en se basant sur des données.

Face à la diversité des outils et des méthodes statistiques, il est cependant possible de donner différentes visions quant aux analyses, aux déductions et donc à la conception d'une nouvelle offre.

Dans le cadre de la mission confiée nous avons, déontologiquement parlant, tenté de garder une neutralité maximale sur l'information délivrée afin de rester au plus proche de la réalité. Des réunions régulières et des contrôles des méthodes d'analyses ont été régulièrement menées dans un souci de qualité des données.

IV.3. Champ d'évolutions possibles

IV.3.1. Des évolutions organisationnelles

Face à l'infobésité, de nombreux dispositifs d'ingénierie documentaire et de gestion des connaissances permettent de mieux agencer et organiser l'information, les documents et les connaissances. L'analyse de l'audience a mis en évidence différents besoins dont les multiples réponses permettraient un meilleur accès aux documents et aux connaissances. Ces nouveaux défis nécessitent une structure de gouvernance de l'information déclinée en ingénierie documentaire, gestion des connaissances, innovation et archivage. En harmonisant les usages suivant les besoins métiers, les organisations accentuent leur spécificité face à la concurrence.

IV.3.1. Une approche processus

Processus métiers à part entière, les déclinaisons de l'ingénierie documentaire et de la gestion des connaissances ne sont pour autant pas le cœur des activités des organisations. Définis en entrant et sortant, l'approche processus permet de cartographier les rôles, les activités et la circulation de l'information, des documents et des connaissances.

En offrant un meilleur accès aux documents importants ou demandés, le processus accompagne le cœur d'activité des organisations contribuant à son excellence opérationnelle.

IV.3.2. Etendre les possibilités techniques

IV.3.2.1. Couplage à un thésaurus

Le classement des documents par le recours au segments répétés en tant qu'approche innovante peut être couplé à d'autres méthodes de classements. Ainsi, en associant l'usage des segments répétés à un thésaurus, l'organisation des termes peut être structurée de diverses façons hiérarchiques et via l'homonymie.

Requêter un thésaurus de thématiques lié aux segments répétés offrirait un mode de classement dynamique (par le thésaurus) et automatiquement adaptable (par l'usage des segments répétés). Les textes seraient indexés dans des thèmes via les segments répétés et les thèmes seraient structurés d'après un thésaurus.

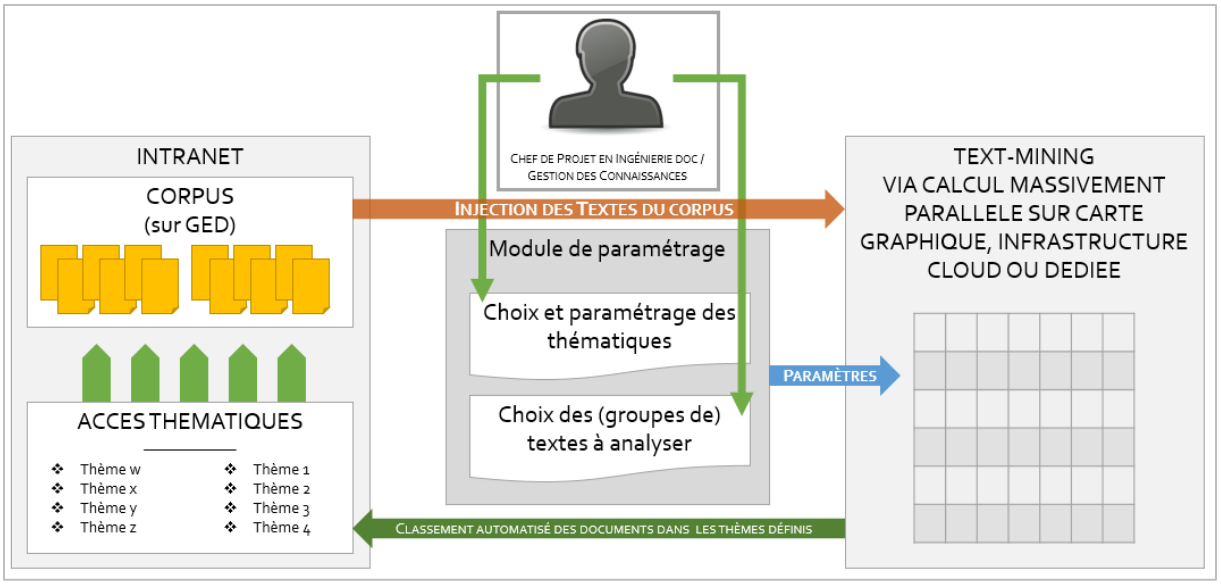
IV.3.2.2. Automatisation de l'indexation des documents dans les thématiques

Bien que la mission de stage ait permis d'automatiser une partie importante du classement des textes dans des thèmes, l'ensemble des actions comporte néanmoins une succession de tâches consommatrices en ressources machines et en temps-homme.

Ces actions sont à réaliser selon une récurrence définie et nécessite un flux de travail convertissant les documents du corpus dans un format exploitable par les outils de text-mining.

Egalement, entre deux analyses complètes, d'éventuels nouveaux documents insérés dans le corpus ne sont pas intégrés dans les thèmes.

En associant diverses technologies actuelles via des développements basés sur l'usage des segments répétés et des infrastructures informatiques dédiées, il serait possible de classer automatiquement les documents dans des thématiques configurées par le chef de projet en ingénierie documentaire et gestion des connaissances sans intervention récurrente.



bb. Proposition de schéma d'architecture permettant de classer automatiquement les documents dans des thématiques

Le schéma ci-dessus comporte trois objets permettant le classement automatique des documents en permanence :

Module de paramétrage dédié au chef de projet

Le module permet de paramétrer les accès thématiques :

- En sélectionnant les groupes de textes issus corpus sur requête.
- Puis en choisissant les segments répétés correspondant au champs lexicaux des thématiques.

Intranet GED

La GED pré-existante comporte le corpus, elle hébergera les accès thématiques configurés et générés

Usage du Calcul massivement parallèle

L'infrastructure de de calcul massivement parallèle, permet d'ajuster en permanence les thématiques en fonction de la vie des documents du corpus.

V. Annexes, Table des matières, Bibliographie

V.1. Annexes

V.1.1. Base de données permettant de concevoir les accès thématiques à partir des segments répétés

V.1.1.1. Introduction

Après intégration du corpus, les outils de text-mining produisent en sortie des données d'analyse nécessitant parfois un retraitement.

Le Tableau des Segments Répétés généré par Lexico classe dans un tableau croisé :

- En colonne : les textes du corpus,
- En ligne : les segments,
- En valeur : la fréquence de répétition des segments par textes.

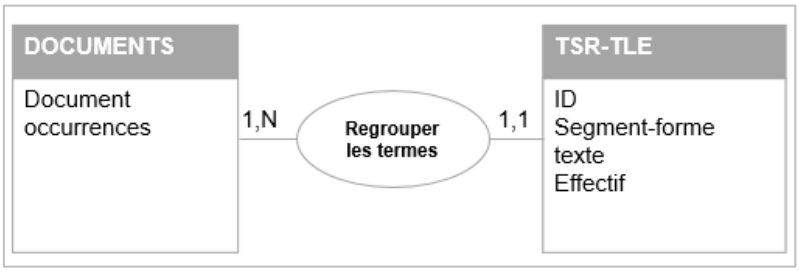
Ces données sont contenues dans un fichier texte.

Après avoir « dépivoté » le tableau en 3 colonnes (segment, texte, fréquence) via l'outil d'extraction Power Query intégré à Power BI et Microsoft Excel, les données sont intégrées dans le SGBD Access.

Lexico Produit un autre tableau comprenant le nombre d'occurrences de mots par textes.

V.1.1.2. Modèles Conceptuel de Données simplifié

Le modèle conceptuel de données ci-dessous représente l'organisation des données extraites depuis le tableau des segments répétés et le tableau lexical entier.

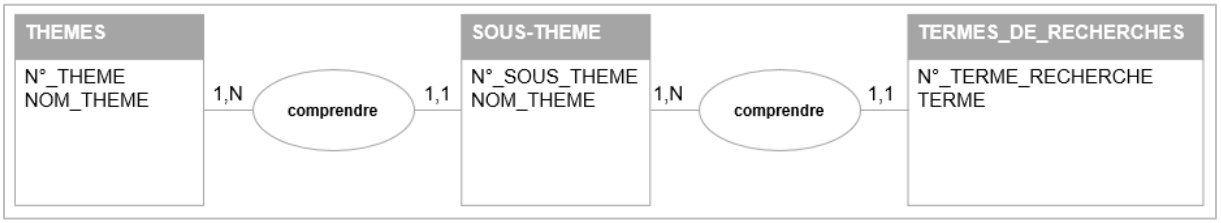


cc. MCD des données de Lexico retraitées

L'entité documents regroupe la liste de documents et le nombre de mots par document.

L'entité TSR-TLE regroupe les formes et les segments et leur effectif par texte.

Le modèle conceptuel ci-dessous représente la structuration des accès thématiques



dd. MCD de la structure des accès thématiques

Un thème regroupe un à plusieurs sous-thèmes comprenant un à plusieurs termes de recherches. Les termes de recherches sont les syntagmes caractéristiques à une sous-thématique.

V.1.1.3. Exemple de requêtes SQL

Un jeu de requêtes liant les termes de recherche aux segments répétés permet d'associer des sous-thèmes à des documents.

Ce jeu permet par exemple de trouver les textes en rapport avec le thème « appétence aux risques » via la recherches de ces termes présents dans la table Termes_de_recherches :

- « appétence au risque »
- « appétence aux risques »
- « matrice d'appétence »

Le jeu comporte certains critères définis dans le mémoire afin d'identifier et classer les documents.

Requête **TSR-TLE-GLOBAL**

```
SELECT [TSR-TLE].*, [TSR-TLE]!effectif/Documents!occurrences*1000 AS
effectif_relatif

FROM Documents INNER JOIN [TSR-TLE] ON Documents.document=[TSR-
TLE].Texte;
```

Cette requête permet de calculer l'effectif relatif d'un terme dans le corpus

Requête **termes_recherchés-segments**

```
SELECT DISTINCT termes_de_recherche.terme, ID, [segment-forme],
texte, effectif, effectif_relatif

FROM [TSR-TLE-GLOBAL] INNER JOIN (SELECT termes_de_recherche.terme
FROM termes_de_recherche GROUP BY termes_de_recherche.terme)  AS
termes_de_recherche ON ([TSR-TLE-GLOBAL].[segment-forme]) Like
termes_de_recherche.terme

GROUP BY ID, [segment-forme], taille, effectif, effectif_relatif,
termes_de_recherche.terme;
```

Cette requête faisant appel à la précédente permet la recherche des syntagmes proposés par sous-thématique dans les segments répétés du corpus afin de classer les documents.

Des requêtes complémentaires affinent et trient les résultats afin de réaliser un classement d'après l'ensemble des règles explicitées dans ce mémoire.

V.2. Table des matières

V.3. Bibliographie

Observatoire des métiers de la banque | les acteurs du système bancaire
observatoire-metiers-banque.fr | consulté en octobre 2018.

Fédération française des banques | Faits et Chiffres N°01 - le secteur bancaire français
fbf.fr | publié en juillet 2018 | consulté en octobre 2018

Le patron d'Orange Bank nous raconte les néobanques | Guerric Poncet
lepoint.fr | publié en aout 2018 | consulté en octobre 2018

Groupe Crédit Agricole | Histoire du Groupe Crédit Agricole
credit-agricole.com | consulté en octobre 2018

Ministère de l'agriculture | les textes fondateurs du monde agricole - Loi du 5 novembre 1894 relative à la création de société de crédit agricole.
agriculture.gouv.fr | consulté en Octobre 2018

Ministère de l'agriculture | loi du 31 Mars 1899 ayant pour but l'institution des Caisses Régionales de Crédit Agricole Mutuel et les encouragements à leur donner ainsi qu'aux sociétés et aux banques locales de crédit agricole mutuel
agriculture.gouv.fr | consulté en Octobre 2018

Conduite de projet informatiques. Développement, analyse et pilotage (4e édition) | Brice Arnaud Guérin
ENI | Parution : aout 2018 | ISBN : 9782409014635
Livre numérique consulté sur eni-training.com en octobre 2018
Chapitre : « *La planification et le chiffrage* » > « *La planification* » > « *1. Les éléments d'un planning* »

La méthode Agile à grande échelle | Andy Noble, Darrell K. Rigby, Jeff Sutherland
hbrfrance.fr | publié le 13 septembre 2018

Manager un équipe projet | Pieric Couteaud Horrut
Support du cours du 8 février 2018

DataViz Quels outils pour quelles datavisualisations ? | Serge Courrier | Page 4
Slideshare.net | mise à jour publiée en septembre 2017 | consulté en Octobre 2018

Analyse de données textuelles
fr.wikipedia.org | consulté en novembre 2018

"Silence Your Phones": Smartphone Notifications Increase Inattention and Hyperactivity Symptoms
Kostadin Kushlev, Jason Proulx, Elizabeth W. Dunn
DOI : http://dx.doi.org/10.1145/2858036.2858359 | publié en mai 2016 – consulté en octobre 2018

Regarder la nature rend plus productif | Nicole Torres
hbrfrance.fr | publié le 10 mars 2018 | consulté en octobre 2018

Visite de l'éco-campus Evergreen, nouveau siège du Crédit Agricole S.A. | Fabrice Mazoir
Blog-emploi.com | publié en juillet 2014 | consulté en octobre 2018

Modélisation conjointe des thématiques et des opinions | Mohamed Dermouche
theses.fr | Page 10 | Thèse soutenue en juin 2015 | consultée en novembre 2018

Ça prend combien de temps de lire un livre ? La durée moyenne de lecture des grands classiques passée au crible | Clémence Jost
Archimag.com | publié en septembre 2014 | consulté en octobre 2018

Retour aux origines de la statistique textuelle : Benzécri et l'école française d'analyse de données
Valérie Beaudouin
Archives-ouvertes.fr | publié en Octobre 2016 | consulté en octobre 2018

Traitement automatique du langage naturel
fr.wikipedia.org | consulté en novembre 2018

NLP, NLU, NLG and how Chatbots work | Anush Fernandes
chatbotslife.com | publié en novembre 2017 | consulté en novembre 2018

LE TAL, FILS PRODIGUE DE L'IA | Gaëlle Recourcé
forum.gfii.fr | publié en décembre 2017 | consulté en novembre 2018

Discours d'entreprise et organisation de l'information Apports de la textométrie dans la construction de référentiels terminologiques adaptables au contexte | Frédéric Erlos
Thèse présentée et soutenue en Novembre 2008 | Page 444
archives-ouvertes.fr | Consulté en décembre 2018

Lexico 3. Outil de statistique textuelle - manuel d'utilisation
Cédric Lamalle, William Martinez, Serge Fleury, André Salem, Béatrice Fracchiolla, Andrea Kuncova, Aude Maisondieu
lexi-co.com | publié en février 2003 | consulté en novembre 2018

Comment préparer l'analyse de textes de sites Web grâce à la lexicométrie et au logiciel Iramuteq ?
Daniel Pélissier | Présence numérique des organisations
presnumorg.hypotheses.org/187 | publié en avril 2016 et mis à jour en mars 2017 | consulté en novembre 2018

Mettre en évidence le temps lexical dans un corpus de grandes dimensions : l'exemple du Parlement européen | Sascha Diwersy, Giancarlo Luxardo | JADT 2016
archives-ouvertes.fr | publié en septembre 2016 | consulté en novembre 2018

Les langages documentaires, principes, histoire et perspectives
Support de cours | Publié en janvier 2018 | Loïc Lebigre

ORLM-255 : Canal+, Netflix, YouTube, Apple, demain la TV !
onrefaitlemac.com | Publié en mars 2017 | Consulté en Décembre 2018

Facettes et systèmes d'information.
Une approche de la classification focalisée sur un besoin de savoir pour agir | Francis Beau
Lavoisier | Les cahiers du numérique | 2017/1 Vol. 13 | pages 115 à 142
Cairn.info | publié en 2017 | consulté en novembre 2018

André Salem, Pratique des segments répétés. Essai de statistique textuelle | Simone Bonnafous
Mots. Les Langages politiques / Année 1988 / n°17 / pages 243-245
persee.fr | consulté en Octobre 2018

Communautés de pratique et performance dans les relations de service, cas des "Front-Office". Quels enseignements pour la GRH ? | Lamine Mebarki
Thèse soutenue en 2011 | consultée en novembre 2018
archives-ouvertes.fr

Classification, codification et appareillage de recherche | S.R. Ranganathan
Unesco.org | consulté en novembre 2018

Une Méthode de classification des énoncés d'un corpus présentée à l'aide d'une application Max Reinert
Les cahiers de l'analyse de données, Tome 15, n°1 (1990) p. 21-36
numdam.org | consulté en novembre 2018

L’analyse de similitude pour modéliser les CHD | Lucie Loubère | JADT 2016
lexicometra.univ-paris3.fr | consulté en novembre 2018

Le Knowledge Management. Un levier de transformation à intégrer
Gonzague Chastenet de Géry
DE BOECK SUP | édition de juin 2018 | ISBN : 978-2-8073-1694-2

La gestion des risques en entreprise: Identifier, comprendre, maîtriser | Jean-David Darsa
Gereso | ISBN : 978-2359534160 | Edition de novembre 2016

MOTEURS D’INDEXATION ET DE RECHERCHE. Environnement client-serveur, Internet et Intranet | Catherine Leloup
Eyrolles | ISBN : 978-2212089769 | édition de juin 1999

Data Mining et Statistique décisionnelle : La science des données | Stéphane Tufféry
Editions Technip | ISBN : 978-2710811800 | édition d'octobre 2017

L'inventaire des segments répétés d'un texte | Pierre Lafon, André Salem
Mots. Les langages du politique / Année 1983 / 6 / pp. 161-177
persee.fr

Approches du temps lexical | André Salem
Mots. Les langages du politique / Année 1988 / 17 / pp. 105-143
persee.fr

Les Langages documentaires. Un panorama, quelques remarques et un essai de bilan
Bruno Menon
A.D.B.S. « Documentaliste-Sciences de l’information » 2007/1 Vol.44 – Pages 18 à 28
Cairn.info | publié en 2007 | consulté en décembre 2018

Thésaurus et informatiques documentaires. Partenaires de toujours ? | Sylvie Dalbin
A.D.B.S. « Documentaliste-Sciences de l’information » 2007/1 Vol.44 – Pages 42 à 55
Cairn.info | publié en 2007 | consulté en décembre 2018

Référentiels terminologiques adaptables au contexte : L’exemple d’un système de recherche d’informations dans une grande entreprise | Frédéric Erlos
lexicometrica.univ-paris3.fr | publié en 2004 | consulté en décembre 2018

Pratique des segments répétés : essai de statistique textuelle | André Salem
Klincksieck Laboratoire Lexicométrie et textes politiques | ISBN : 978-2-252-02549-9 | 1987