Journées de méthodologie statistique de l'Insee 2022

Identification de collectivités territoriales en utilisant un moteur de recherche indexé à partir d'RMéS

Théo LEROY (*)

(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale

theo.leroy@insee.fr

Mots-clés. (6 maximum) : Code officiel géographique, RDF, moteur de recherche, ElasticSearch

Domaines. Codification automatique, Open data

Résumé

Le besoin de coder des libellés géographiques dans le Code Officiel Géographique (COG) revient dans de nombreuses productions de l'Insee (le recensement de la population, le traitement des déclarations sociales nominatives ou la constitution de Fidéli). Dans ces cas, la codification est assurée par l'outil mutualisé Sicore. Cependant, chaque année, la mise en oeuvre est particulièrement lourde car elle demande une actualisation en partie manuelle des bases de connaissances Sicore à partir du COG. De plus, Sicore ne répond pas au besoin grandissant en matière d'autocomplétion car ne permet de coder que sur un libellé complet.

Dans ce contexte, cet article présente une alternative aux environnements Sicore pour la géographie reposant sur le moteur de recherche ElasticSearch. Les moteurs de recherche permettent de retrouver des documents selon un critère de similarité défini en fonction du besoin. Ils se prêtent particulièrement bien pour ces cas d'usage variés. En effet, il est souhaitable, par exemple, de plutôt valoriser la concordance sur les premiers caractères dans le cas de l'autocomplétion contre tous dans le cas d'un libellé complet.

Les données du COG ont été indexées dans le moteur de recherche directement à partir du graphe RDF du COG. Un raccordement automatique entre le système de codage (le moteur de recherche) et la base de vérité (le COG) existe et assure une fraîcheur dans les données mobilisées par le système de codage. Néanmoins, l'index, produit exclusivement à partir du COG, peut être perçu comme insuffisant. En pratique, on dispose parfois de codes postaux en plus du libellé de la commune sur lequel il serait possible de s'appuyer. Par ailleurs, on trouve régulièrement des libellés de lieux-dits absents du COG à la place d'un vrai nom de commune dans la source à coder. L'index a donc été enrichi en ajoutant des attributs provenant d'autres sources en interrogeant leurs services web (Laposte pour les codes postaux et OpenStreetMap pour les lieux-dits ou les communes frontalières).

Pour coder en s'appuyant sur l'index précédemment décrit, une méthode de recherche par relâchement de contraintes a été mise en oeuvre. Une requête d'abord très stricte sur le libellé et le code département ou postal est effectuée. Si l'entité géographique n'a pas pu être identifiée, plusieurs passes avec des critères de similarité entre l'input à coder et les documents contenus dans l'index de plus en plus souples sont exécutées. Cette approche donne des performances (taux de libellés codés automatiquement et taux de bien codés) légèrement meilleures que Sicore Commune sur une base de données de communes annotée à la main extraite de la base tous salariés 2017.

Abstract

In the official statistical service, the coding of towns is often necessary to process data and make studies. The method currently used is not very robust and relies on a hand-made dictionary. This article puts forward a new method directly linked to the INSEE metadata repository and on usual scores of the information retrieval theory (Okapi BM25).

1 Introduction

1.1 Présentation de l'existant

Une codification des communes dans le code officiel géographique (COG) à partir d'un code postal ou d'un code de département et d'un libellé de commune est réalisée dans un certain nombre de productions de l'Insee, par exemple pour les besoins du recensement de la population ou pour la créations des bases Tous salariés. Ces opérations reposent sur l'outil Sicore [1] et une base de connaissances dédiée. Cette base de connaissances est principalement élaborée autour d'un index qui permet d'associer à chaque code commune un ou plusieurs libellés.

Code dans le COG	Libellé (Code de département + nom de la commune)
17295	17 REAUX
17295	17 REAUX SUR TREFLE
17296	17 RETAUD
17297	17 RIVEDOUX PLAGE
17298	17 RIOUX
17299	17 CANAL DES SOEURS
17299	17 CENTRE COMMERCIAL MARTROU
17299	17 CORDERIE ROYALE
17299	17 L ARSENAL
17299	17 LES 4 ANES
17299	17 PONT NEUF
17299	17 ROCHEFORT
17299	17 ROCHEFORT MER

TABLE 1 – Extrait de l'index des communes Sicore des codes communes 17295, 17296, 17297, 17298 et 17299

Cet index est plus riche qu'un simple dictionnaire des communes et comprend près de 50 000 éléments (contre 14 dans l'extrait de la table 1) alors qu'on compte seulement 35 000 communes dans le COG. En effet, pour un même code, on trouve parfois des formulations différentes pour la même entité. Par exemple, pour le code `17295`, la base de connaissances Sicore contient le libellé `REAUX SUR TREFLE` (nom officiel) et un nom plus simple d'usage `REAUX`. De plus, des zones géographiques comprises dans la communes comme des lieux dits, des de quartiers ou des zones commerciales sont aussi parfois référencées.

La mise à jour de cette base de connaissances est essentielle pour disposer d'une codification automatique efficace et fiable car toute la codification repose sur elle. Il faut parfois ajouter ou supprimer des libellés ou bien gérer les mouvements de communes liés aux fusions ou partitions d'entités du COG. Ces opérations sont en partie réalisées manuellement et sont donc coûteuses et des sources d'erreurs ou d'oublis.

Lors de la codification d'un libellé de commune avec Sicore, le libellé est d'abord normalisé puis ce libellé transformé est confronté à l'index où les libelles ont également été normalisés au préalable. L'étape de normalisation est relativement simple pour la variable commune. Elle consiste principalement à traiter des caractères spéciaux comme le tiret ou l'apostrophe et homogénéiser certaine mots. Par exemple, les mots « SAINT », « SAINTE », « ST » ou « STE » sont considérés comme identiques et normalisés par le même symbole. La confrontation entre l'index et le libellé à coder est plutôt stricte et la codification échoue en général dès la première faute d'orthographe.

1.2 Pistes d'amélioration de l'existant et objectifs de l'expérimentation conduite

L'objectif des travaux présentés ici est de concurrencer la codification automatique via Sicore Commune autour de deux axes.

Le premier consiste à s'orienter vers un index construit automatiquement qui permettrait d'être plus riche que celui existant sans demander d'intervention humaine. La brique principale de cet index serait sur le référentiel des métadonnées statistiques (RMéS) [2] qui est la source de vérité dans la diffusion de COG. Cette utilisation des métadonnées de manière active permettrait de construire un système de codification où l'information mobilisée pour coder est fraîche et cohérente avec la diffusion.

Toutefois, l'index n'est pas la seule piste d'amélioration du système actuel. La phase de reconnaissance d'un libellé au sein de l'index pourrait être perfectionnée en testant des alternatives à celle mise en oeuvre dans Sicore. EN effet, cette dernière s'apparente à un appariement strict sur les libellés normalisés. Dans ces travaux, des méthodes exploitant des scores de similarité (Levenshtein, tf-idf, Okapi BM25... [3]) ont été testées afin d'être plus robuste que Sicore sur d'éventuelles fautes d'orthographe.

2 Acquisition des données et indexation

Une base de données à l'image de l'index Sicore a été construite en mobilisant plusieurs sources. Elle est alimentée grâce à programme de type « crawler » qui permet d'acquérir des données pour initialiser ou mettre à jour périodiquement la base de données. Pour répondre au problème d'information textuel, le système de base de données retenu dans le cadre de cette expérimentation est un moteur de recherche appelé ElasticSearch [4] qui offre de nombreuses fonctionnalités pour l'indexation et la recherche d'information.

2.1 Origine des données

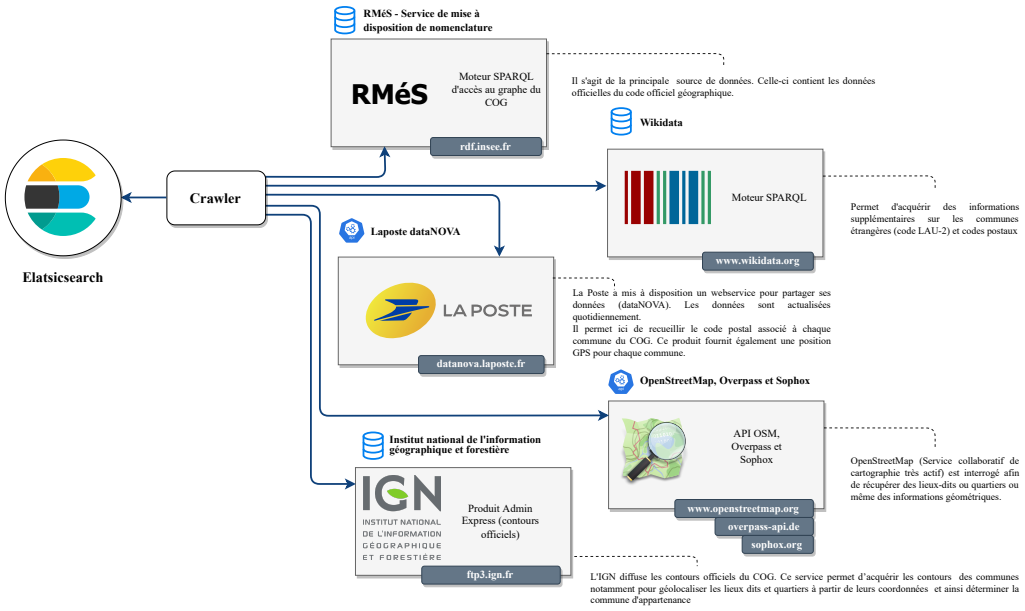

FIGURE 1 – Présentation des différentes sources de données mobilisées

A l'inverse, des moteurs de recherche usuels qui parcourent l'ensemble du web, l'information indexée ici provient d'un nombre limité de sources choisies au préalable. Les services interrogés sont décrits dans la figure 1 L'information officielle du COG accessible en *open data* est la principale source de données. Cependant, celle-ci ne recouvre pas tout le contenu de l'index Sicore. Pour tenter d'égaler ou dépasser l'index Sicore, des attributs supplémentaires sont récupérés par ailleurs dans des sources officielles comme La Poste ou l'IGN ou des services collaboratifs comme Wikidata (base de connaissances structurée) ou OpenStreetMap (projet de cartographie) qui contiennent d'abondantes données mais de qualité moindre (voir figure 2). Au cours de cette étape d'acquisition de données, certaines caractéristiques non présentes dans l'index historique Sicore comme les codes postaux ont été collectées. Cette information plus précise que le code département pourrait être pertinente lorsque les données à coder contiennent de tels codes en plus du libellé de la commune. Les communes étrangères de certains pays frontaliers ont également été intégrées dans le but d'améliorer l'efficacité de la codification automatique.

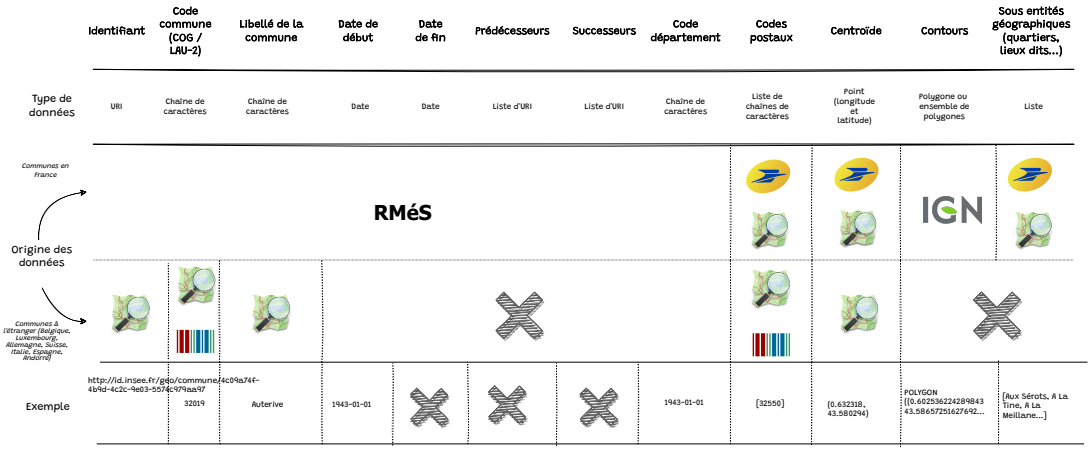

FIGURE 2 – Structure et origine des données acquises

2.2 Indexation

Les données sont stockées dans le moteur de recherche sous forme de documents (ici des communes). Chaque document possède des champs c'est à dire des variables qui caractérisent la commune. Pour chaque champ, on peut définir la phase de pré-traitement des données réalisée lors de l'indexation mais aussi recherche au travers de ce champ. Cette phase préalable comprend plusieurs étapes. Elle intègre une dimension de normalisation comme la possibilité de définir des remplacements ou suppressions de certains caractères ou groupe de caractères. Cette fonctionnalité a été principalement utilisée pour reproduire la normalisation opérée dans l'outil de codification Sicore (ST=SAINT, PT=PONT, HT=HAUT...). De plus, cette phase de pré-traitement comprend une étape de tokenisation. Elle consiste à découper d'information en plus petits éléments. Pour du texte, les tokens choisis sont souvent des mots ou des ngrams (c'est à dire des séquences de taille fixe de caractères consécutifs). Ce choix de découpage est très important car, au moment de la recherche, le score de pertinence d'un document fait généralement intervenir le nombre de tokens en commun entre la chaîne en entrée et le document. Plusieurs champ ont été définis à partir des informations collectées. Ils sont décrits dans la table 2.

Variable d'origine	Normalisation	Tokenisation
Nom de la commune	Semblable à Sicore*	Mot
Nom de la commune	Semblable à Sicore + suppression des précisions sur le lieu dans le nom (par exemple "SUR LOIRE, "EN BUGEY" ...)	Mot
Nom de la commune	Semblable à Sicore	Ngram de 2 à 3 caractères
Nom de la commune	Semblable à Sicore	Mot avec transformation des token dans un langage phonétique (Beider Morse)
Libellés de sous zones géographiques	Semblable à Sicore	Mot
Libellés de sous zones géographiques	Semblable à Sicore	Ngram de 2 à 3 caractères
Libellés de sous zones géographiques	Semblable à Sicore	Mot avec transformation des tokens dans un langage phonétique (Beider Morse)
Codes postaux	Conservation des caractères numériques uniquement	Mot
Code de département	Conservation des caractères numériques uniquement	Mot

* : traitement des abréviations, suppression des mots vide de sens, uniformisation de la casse, suppression de l'accentuation

TABLE 2 – Champs définis dans l'index

3 Méthodologie de la recherche

Une méthode de recherche au travers de l'index a été élaborée pour la codification automatique. La recherche se déroule sous forme d'une suite de requêtes relâchant de plus en plus les contraintes sur la conformité entre l'information à coder et celle contenue dans l'index. Elle est réalisée au plus en 4 étapes (voir figure 3) et conduit à un code ou aucun si la dernière requête n'a renvoyé aucun résultats ou des résultats trop incertains (l'écart entre les deux scores les plus élevés est trop faible).

Les trois premières requêtes imposent une ressemblance forte entre le libellé à coder et le contenu de l'index. La distance de Levenshtein qui comptabilise le nombre de transformations élémentaires entre deux chaînes (ajout, suppression ou substitution d'un caractère) ne peut excéder la valeur de 1. La dernière requête se base sur un score de pertince appelé Okapi BM25 qui s'inspire du score tf-idf. Il est construit également à partir de la quantité tf (term frequency). Plus un token du libellé à rechercher apparaît de nombreuse fois dans un document et plus le score de ce document est élevé. La quantité idf (inverse document frequency) est l'inverse de la proportion des documents de l'index qui contiennent ce token. Cela permet de mesurer l'importance d'un terme dans l'ensemble de l'index et valoriser les tokens très spécifiques qui se trouvent uniquement au sein d'un nombre restreint de documents. Le score Okapi BM25 permet en plus de ternir compte de la longueur du champ d'un document. Si un token de la requête apparaît dans des documents, on favorise les documents les plus courts car ce token à plus d'importance au sein de ces documents.

A l'issue de la recherche, si la commune trouvée n'est pas dans le millésime du COG souhaitée, des calculs de successeurs ou prédécesseurs sont réalisés pour coder à la date désirée.

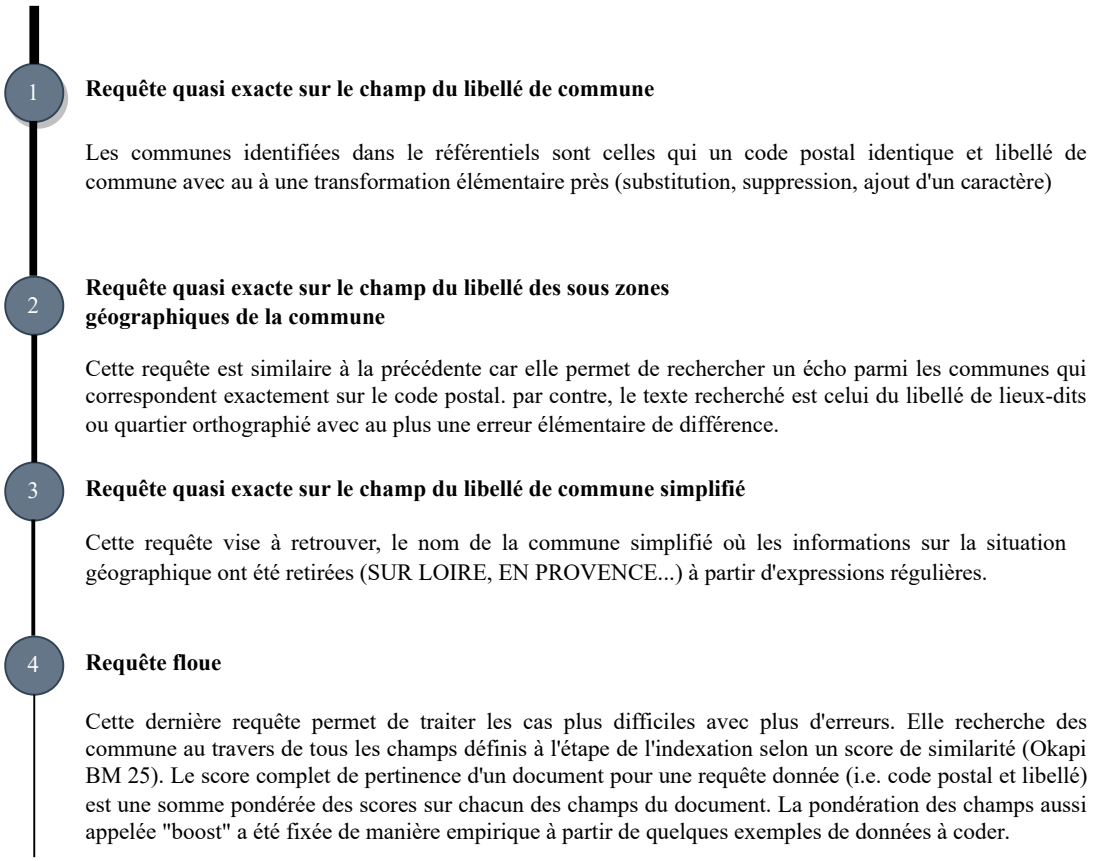

FIGURE 3 – Succession de requêtes réalisées pour la codification automatique à partir de l'index

4 Résultats

Pour étudier les performances de cette solution deux indicateurs ont été analysés : le taux de de libellé codés automatiquement qui permet de mesurer l'efficacité de la méthode et le taux d'observations bien codées c'est à dire la fiabilité du codage automatique. Cette évaluation a été conduite sur la commune de résidence du salarié dans la base Tous salariés 2017 contenant plus de 27 millions d'observations. Toutefois, la commune de résidence est renseignée de manière plutôt homogène. On compte moins de 230 000 valeurs différentes du couple code postal et libellé de commune non-normalisé.

		Nombre de cas uniques (code postal, libellé) codés	Nombre d'observations codées
Sicore	Fréquence	120 000	26 576 000
	Pourcentage	52,95	98,29
Méthode concurrente	Fréquence	225 000	27 032 000
	Pourcentage	99,13	99,98

TABLE 3 – Efficacité de la méthode de codification automatique sur la codification de la commune de résidence de la base Tous salariés 2017

La méthode testée à partir du moteur de recherche ElasticSearch est légèrement plus performante avec plus d'un point de codification automatique supplémentaire. Cette efficacité accrue s'explique en partie par l'index plus riche (notamment sur les quelques communes étrangères)

mais surtout par la robustesse de la codification offerte par cette méthode élaborée autour de score de similarité en dernière intention.

Code postal	Libellé	Nombre d'observations	Codage méthode concurrente	Codage Sicore
94400	94081 VITRY SUR SEINE	87	94081	
94400	BVITRY SUR SEINE	1	94081	
94400	VIRTY SUR SEINE	32	94081	
94400	VIRY SUR SEINE	11	94081	
94400	VITRY SUR SEINE	48 903	94081	94081
94400	VITRYSUR SEINE	14	94081	
94400	VITRYSURSCNE	1	94081	
94400	VITRYSURSEINE	519	94081	

TABLE 4 – Exemple de codification sur des communes de résidence de la base Tous Salariés 2017 avec le code postal 94400

Le système de codification existant est très difficile à concurrencer car il est très bon sur les cas les plus fréquents. Ces exemples sont, en effet, contenus dans l'index. Cependant, cet outil n'est pas très robuste pour gérer des cas qui s'en écartent légèrement et dans ces circonstances la méthode alternative se démarque. Une extraction des données avec le code postal 94400 illustre ce défaut dans la table 4. Sicore serait ainsi très peu intéressant en cas de reprise manuelle des communes car plus de 100 000 couples de code postal et libellé devraient être repris contre moins de 2 000 avec la méthode construite ici à partir d'un moteur de recherche.

Pour mesurer, la qualité, toutes les observations n'ont pas pu être annotées pour déterminer la valeur de vérité (ground truth). Un échantillon de 10 000 observations a été tiré (dans la base des cas uniques) en stratifiant selon les 4 cas suivants :

— Codage identique entre Sicore et la méthode alternative
— Codage dans les deux méthodes mais non concordant
— Codage uniquement avec la méthode alternative
— Codage uniquement par Sicore

Un taux de sondage très faible a été choisi pour la première strate car on pouvait s'attendre à peu d'erreur et par conséquent une estimation de la variance de la proportion de bien coder très faible dans cette strate. Au sein de chaque strate, les tâches à coder ont été échantillonnées avec des probabilité d'inclusion proportionnelles au nombre d'occurrence grâce à un tirage systématique. En effet, la variable d'intérêt (le nombre d'observation bien codée) est proportionnel au nombre d'occurrence (car cette variable d'intérêt vaut 0 si le codage est erroné ou le nombre d'occurrence sinon). L'estimation de la proportion de commune de résidence bien codées au travers de ces deux méthodes est donnée dans le tableau 5. Les deux méthodes sont équivalentes avec un léger avantage pour la méthode proposée dans cet article.

	Estimation de l'intervalle de confiance à 95 % de la précision (% du taux de bien codés)	
	Sur les cas uniques (code postal, libellé) codés	Sur l'ensemble des données codées
Sicore	[85,3 ; 92,9]	[99,4 ; 99,5]
Méthode concurrente	[85,5 ; 91,5]	[99,7 ; 99,7]

TABLE 5 – Fiabilité de la méthode de codification automatique sur la codification de la commune de résidence de la base Tous salariés 2017

5 Conclusion

Cette expérimentation de méthode alternative à Sicore pour la codification des communes ne semble pas dégrader la qualité très élevée de la codification. La technique mise en oeuvre conduit à une efficacité et une fiabilité au moins aussi bonne sans engendrer de temps de calcul supplémentaire sur le jeu de données testé. Bien que la méthode n'aait pas été défini à partir de l'échantillon de données annotés, une évaluation sur d'autres sources pourrait confirmer ou non les performances présentées dans cet article.

En dehors des questions de qualité, cette solution se distingue sur des questions pratiques et fonctionnelles. La mise a jour est plus simple car automatique et provient directement du code officiel géographique. Par ailleurs, l'usage d'un moteur d'indexation recherche permet d'offrir des services supplémentaires et il est possible d'indexer les données pour offrir de la recherche sur un libellé en cours de saisie (autocomplétion).

Références

[1] RIVIERE Pascal. Sicore, un outil et une méthode pour le chiffrement automatique à l'Insee", Courrier des statistiques, n°74 - août 1995

[2] BONNANS Dominique. RMéS, le référentiel de métadonnées statistiques de l'Insee, Courrier des statistiques N2 - 2019

[3] MANNING Christopher, RAGHAVAN Prabhakar, SCHUTZE Hinrich. An Introduction to Information Retrieval, Cambridge University Press, 2009.

[4] GORMLEY Clinton, TONG Zachary. Elasticsearch : The Definitive Guide : A Distributed Real-Time Search and Analytics Engine, O'Reilly Media, 2015