Organisation de Coopération et de Développement Économiques

DAF/COMP/WP3/WD(2022)30

Non classifié

Français - Or. Français

3 novembre 2022

DIRECTION DES AFFAIRES FINANCIÈRES ET DES ENTREPRISES
COMITÉ DE LA CONCURRENCE

Groupe de travail n° 3 sur la coopération et l'application de la loi

Les outils de filtrage des données dans les enquêtes de concurrence – Contribution de la France

28 novembre 2022

Ce document est une contribution écrite soumise par la France au titre de la session 3 de la 136ème réunion du Groupe de Travail 3 tenue le 28 novembre 2022.

D'autres documents relatifs à cette discussion sont disponibles sur :.
https://www.oecd.org/daf/competition/data-screening-tools-for-competition-investigations.htm

Mme Despina PACHNOU
[Email: Despina.PACHNOU@oecd.org].

JT03506579

France

Introduction

1. Si la numérisation toujours plus poussée du monde dans lequel nous vivons induit des transformations profondes de l'économie (émergence de plateformes numériques, enjeux concurrentiels complexes à appréhender, etc…), l'innovation technologique sur laquelle elle est fondée peut aussi bénéficier aux autorités de concurrence et venir renforcer et améliorer les outils dont elles disposent pour mener à bien leurs missions.

2. En 2020, l'Autorité de la concurrence (l'Autorité) a décidé de créer, au sein des services d'instruction, un service de l'économie numérique (SEN) visant à renforcer ses moyens en matière numérique. Parmi les objectifs de ce service, actuellement composé de quatre personnes, dont deux *data scientists*, figure celui de développer de nouveaux outils numériques à même d'aider les services d'instruction dans leur travail d'investigation.

3. Parmi ces nouveaux outils, les outils de filtrage des données (*screening tools)* peuvent notamment apporter une aide précieuse aux rapporteurs dans la détection de pratiques anticoncurrentielles : ententes sur les prix, prix imposés, répartition géographique de marchés sont parmi les pratiques pour lesquelles ces outils peuvent s'avérer très efficaces. Ces outils peuvent également servir dans l'automatisation de certaines tâches de surveillance du marché.

4. La mise en place d'outils de filtrage des données s'articule généralement autour de deux pôles distincts mais interconnectés : un outil de collecte de données et un outil de visualisation des données et des indicateurs pertinents qui y sont associés.

5. La présente contribution détaillera tout d'abord les deux principales méthodes de collecte des données (I), puis présentera les principaux éléments que, selon l'Autorité, l'outil de visualisation doit comporter (II), et finira par illustrer l'exemple d'un outil actuellement en cours de développement par le SEN au sein de l'Autorité.

1. Collecte de données

6. Les outils de filtrage des données servent généralement à placer un marché sous surveillance et/ou à essayer de détecter des pratiques anticoncurrentielles (concertation sur les prix, prix imposés, répartition géographique de marchés, etc.) sans que les entités concernées n'en soient informées, au moins dans un premier temps, et pendant une certaine durée.

7. Pour ces raisons, ces *screening tools* sont généralement alimentés par des sources de données externes à l'Autorité, et non par des données qui auraient été collectées par l'intermédiaire des pouvoirs d'enquête de l'Autorité.

8. Deux options principales existent pour l'alimentation en données : les API (1.1) et les méthodes de *scraping* (1.2).

1.1. Les API

9. Les API (*application programmation interface* ou interface de programmation informatique) permettent d'accéder à de nombreuses données dans le cadre d'un accès cadré.

10. Au travers d'une API, le propriétaire d'une base de données peut donner accès à des tiers de manière fine. Il peut déterminer les données qui seront accessibles, les bénéficiaires de l'information, et peut même graduer l'accès en définissant différents types de profils.

11. Les API permettent donc un accès sécurisé et simple à certaines données. L'architecture technique est prise en charge par celui qui met à disposition l'API et toute personne intéressée par ces données peut y avoir accès en respectant le protocole qui aura été déterminé. Par ailleurs, les API permettent souvent un accès à l'ensemble de l'historique des données, ce qui permet aux *screening tools* utilisant ce type de données de pouvoir faire des analyses rétrospectives.

12. En France, un principe d'ouverture des données publiques a été décidé par une circulaire du Premier ministre du 26 mai 2011[1] ; de nombreuses données sont disponibles sur le site internet « data.gouv.fr » et un ensemble d'API est disponible sur le site internet « api.gouv.fr ».

13. En conclusion, les API sont une solution idéale pour des outils de filtrage des données envisagés sur le long terme. Il est cependant utile de noter que, lors de l'utilisation d'une API, l'utilisateur est très souvent authentifié, cette solution ne pouvant dès lors être envisagée si l'utilisation du *screening tool* nécessite un certain anonymat. De plus, toutes les données ne sont pas forcément disponibles via une API et la méthode du *scraping* peut dans cas de figure permettre de récupérer des données accessibles en ligne.

1.2. Le Scraping

14. Le *scraping* est une technique d'extraction automatisée du contenu de sites internet ou d'applications, web ou mobile, via un programme informatique. Les données récupérées sont ensuite généralement structurées dans une base de données.

15. Pour chaque utilisation, un travail de repérage doit être réalisé en amont afin d'identifier l'emplacement des informations pertinentes et de déterminer la fréquence de récupération des données souhaitée (heure par heure, journalière, hebdomadaire, etc.).

16. La mise en place technique d'un *scraping* nécessite une certaine technicité, mais de nombreuses librairies en ligne permettent de faciliter ce travail. Sous le langage de programmation « Python », on peut notamment citer « Scrapy » ou « Beautiful soup ».

17. Par ailleurs, chaque solution de *scraping* mise en œuvre est adaptée à l'architecture d'une page à un instant donné ; lorsque cette dernière est mise à jour, l'outil peut nécessiter une réadaptation.

18. Pour l'ensemble de ces raisons, les méthodes de *scraping* ne sont pas optimales pour les projets de long terme, car elles nécessitent un monitoring humain de la partie technique bien plus important que si l'on utilisait une API. Enfin, cette méthode ne permet pas d'analyses rétrospectives dans la mesure où l'accès aux données n'a lieu qu'à partir du moment où la méthode de *scraping* est mise en place.

[1] Circulaire du 26 mai 2011 relative à la création du portail unique des informations publiques de l'État « data.gouv.fr » par la mission « Etalab » et l'application des dispositions régissant le droit de réutilisation des informations publiques.

2. Outil de visualisation et d'analyse

19. Une fois les données collectées, il est nécessaire de les valoriser à travers leur visualisation et leur analyse. Il convient d'éviter à tout prix un effet « boîte noire » où les données ne seraient analysées qu'une fois la période de collecte terminée.

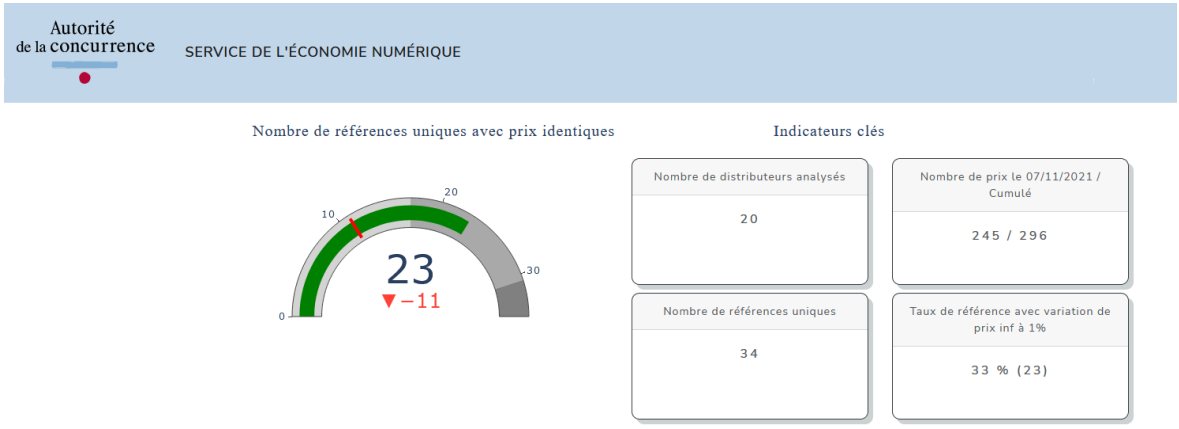
20. En effet, les outils actuels permettent de suivre et de calculer différents indicateurs en temps réel, qui s'actualisent chaque jour pendant toute la période de monitoring. Les indicateurs les plus classiques que l'on peut citer sont par exemple le taux d'alignement de prix (si l'on cherche à détecter une entente sur les prix) ou encore des indicateurs géographiques avec des cartes associées (si l'on veut mettre en évidence une répartition géographique de marchés).

21. L'identification, la création et la mise en place des indicateurs appropriés sont des éléments cruciaux pour la bonne réussite d'un outil de filtrage des données. Ces indicateurs sont propres à chaque dossier, et par conséquent, un bon outil de visualisation se met en place en coordination avec les rapporteurs en charge du dossier concerné, afin de pouvoir sélectionner, créer et présenter les indicateurs les plus pertinents.

22. À l'Autorité, les outils de visualisation utilisés sont réalisés sous « Dash » (*Framework Python open source*). Ils se présentent sous la forme de tableaux de bord et présentent généralement deux types d'analyses, accessibles chacune depuis un onglet spécifique.

23. La première typologie d'analyse présente les statistiques générales et les indicateurs globaux, un exemple étant illustré en Figure 1.

Fig 1 : Exemple d'indicateurs généraux issus d'un projet de démonstration réalisé par le SEN



24. La deuxième typologie d'analyse présente des analyses individuelles, généralement aptes à suivre l'évolution des données collectées dans le temps, sur la base d'une segmentation plus fine à travers une barre de recherche. À titre d'exemple, il peut s'agir d'une page sur laquelle il est possible de suivre l'évolution des prix d'un produit dans le temps, ou bien d'une page qui permet de retrouver toutes les caractéristiques d'un marché public passé.

3. Exemple d'outils de filtrage des données à l'Autorité : détection de soumissions concertées dans les marchés publics

25. Le SEN a lancé en fin d'année 2020 un projet de détection d'ententes dans les marchés publics en France en utilisant les possibilités offertes par la *data science*. L'idée au cœur de ce projet est de mettre à profit les nombreuses données disponibles concernant les marchés publics en France et de les combiner avec d'autres sources de données publiques, afin de créer des indicateurs permettant de détecter des anomalies dans les soumissions aux procédures de marchés publics.

26. Dans un deuxième temps, ces anomalies devront être analysées par les services d'instruction pour confirmer ou non l'existence d'une entente anticoncurrentielle.

27. La mise en place de ces outils n'est pas une tâche aisée et de nombreuses autres autorités de concurrence se sont lancées dans des projets similaires depuis plusieurs années. Il s'agit ainsi d'un projet de long terme dont les résultats ne sont pas attendus avant plusieurs années.

28. En France, deux sources de données publiques concernant la commande publique sont disponibles, à savoir les données essentielles de la commande publique (DECP[2]) et le bulletin officiel des annonces des marchés publics (BOAMP[3]).

29. S'agissant des DECP, les données concernant les marchés publics sont disponibles depuis le 1er octobre 2018 et représentent environ 100 000 marchés publics par an. Pour ce qui est du BOAMP, les données concernant les marchés publics sont disponibles depuis le 1er janvier 2015 et représentent environ 160 000 marchés publics par an.

30. Dans le cadre du projet mené par le SEN, il nous a semblé utile de combiner les deux sources de données mentionnées à une troisième source, complémentaire par rapport aux deux premières : le registre national du commerce et des sociétés (RNCS[4]), qui centralise l'ensemble des informations sur les sociétés immatriculées en France.

31. Au stade actuel du projet, la connexion aux bases de données évoquées ci-dessus est réalisée et opérationnelle (Figure 2). Le SEN s'attache désormais à concevoir les différents indicateurs pertinents qui devront permettre d'identifier les marchés publics présentant des anomalies. Il s'agit d'un travail complexe, souvent chronophage et qui nécessite une certaine ingéniosité. En effet, il est nécessaire de trouver des solutions pour palier à l'absence de certaines informations, par exemple en lien avec les soumissions perdantes. Il n'est par ailleurs pas possible à l'heure actuelle de faire appel à des méthodes de *machine learning,* faute de données labellisées suffisantes.

32. Au vu de ces contraintes, le projet ne s'attache donc pas à avoir un fort taux de détection par marché public (on accepte une présence importante de faux négatifs), mais espère compenser cet écueil par la quantité de données disponibles. Ainsi, même avec un faible taux de détection, il est possible d'aboutir à la détection de plusieurs cas par an lorsque ce taux est appliqué à plus de 100 000 marchés publics par an. Il convient donc, dans un premier temps, de privilégier la quantité à la qualité.

[2] Données essentielles de la commande publique - fichiers consolidés (DECP) - data.gouv.fr

[3] Accueil | Pages — boamp.fr

[4] Qu'est-ce que le Registre national du commerce et des sociétés ? | 1 INPI.fr

33. Il est par ailleurs possible d'espérer que la détection des premiers cas d'entente anticoncurrentielle permette au SEN d'améliorer progressivement l'outil de détection, conduisant ainsi à détecter davantage de cas à l'avenir. C'est en tout cas l'objectif souhaité.

Fig 2: Projet d'architecture technique à code source ouvert