Interview de Deivyd Vélasquez, doctorant à Bordeaux Sciences Economiques – 11/2022

0'00 : Présentations de Karine Onfroy et Deivyd Vélasquez

Bienvenue,
Je suis Karine Onfroy, statisticienne à Bordeaux Sciences Economiques (BSE pour les initiés), une Unité Mixte de Recherche du CNRS et de l'Université de Bordeaux.
Je ne sais pas pour vous mais le retour d'expérience, c'est toujours le moment qui, soit me rassure et j'avance ou, à l'inverse qui me fait abandonner toutes initiatives.
Ce n'est pas grave, c'est aussi à ça que ça sert. Alors je me suis dit que ce serait une bonne idée d'aller à la rencontre des doctorants de mon laboratoire qui ont osé lever le secret statistique …
Toc, toc, toc,
Bonjour Deivyd, je peux te déranger ?
Je réalise le podcast sur les démarches pour accéder à des données confidentielles via le CASD.
On commence l'interview et je te laisse te présenter.

Bonjour,
Je suis Deivyd Vélasquez, je suis doctorant au BSE, je fais une thèse en sciences économiques financée par le Conseil régional de la Nouvelle Aquitaine. Mon sujet de thèse est « Analyser les dynamiques en innovation dans la région Nouvelle Aquitaine à partir d'un cadre théorique et d'un cadre pratique ».

1'10 : Sur quelles données travaille Deivyd ?

Aujourd'hui tu travailles sur des données confidentielles. Ton environnement de travail se présente sous la forme d'une box sécurisée, tu ne peux pas imprimer tes sorties graphiques, on ne peut pas regarder ton écran … bref, Deivyd pourquoi tout ce mystère ? C'est quoi exactement tes données ?

Ces données ne sont pas agrégées comme habituellement on les trouve sur Internet. Il y a une finesse presque nominative à l'échelon territorial ou par secteur.
Mon travail jusqu'à maintenant se fait avec deux producteurs : l'INSEE et le Ministère de l'enseignement supérieur. Le niveau de détail, c'est la finesse de tous les acteurs de la Nouvelle Aquitaine qui développent des activités en innovation et en recherche.

Donc tu as eu besoin de cette finesse d'observation, tu as le nom des entreprises en fait ?

Exact, j'ai le nom, mais aussi tous les secteurs et toute l'information que l'on ne trouve pas facilement sous les bases publiques ou sous Internet. Donc, cette information est très utile pour mon travail pour mieux comprendre ce qui se passe au niveau de l'innovation dans le territoire mais au niveau de chaque acteur.

Je vais utiliser ces deux bases de données mais je vais les apparier avec d'autres bases de données plus connues au niveau des brevets, des publications scientifiques, des projets de recherche.

Donc, tu travailles au niveau d'un acteur, une observation pour toi, c'est un acteur.

Exactement.

2'55 : Avec quels logiciels travaille Deivyd ?

Et tu as choisi de travailler sur quels logiciels ?

Les logiciels de base pour les études statistiques : Excel, R et Python.

3'09 : Pourquoi Deivyd a-t-il choisi les données utilisées par le CASD ?

Si l'on revient un petit peu en arrière, comment en es-tu venu aux données utilisées sur le CASD ? Tu avais une idée des sources dont tu avais besoin au départ ?

Au départ, non parce que mon sujet est très vaste en innovation et recherche.
Donc, je suis allé sur le site du CASD pour regarder quelles bases de données étaient disponibles et surtout quelles étaient les variables disponibles. C'est très bien construit, j'ai cherché directement tous les sujets autour de mon sujet de thèse, toutes les bases de données autour de mon sujet de thèse. Et puis j'ai choisi les deux bases que j'ai mentionnées avant.

Donc, tu as bien regardé avec attention ce qui était proposé, après tu as fait ton choix.

Exactement.

C'était à quel moment de ta thèse ?

Officiellement j'ai commencé à la moitié de ma deuxième année, c'était plus ou moins en décembre.

4'12 : Quand et comment Deivyd a-t-il reçu les données ?

Les données tu les as reçues quand exactement ?

Officiellement, après avoir fait toutes les démarches au CASD, c'était en juin de cette année. Donc, j'ai commencé la démarche en décembre, c'est plus ou moins 6 mois après.

Et durant ce processus entre décembre et juin, est-ce que tu as effectué des changements par rapport à ton idée initiale ?

Oui, tout à fait.

Principalement parce que, comme il y a beaucoup de producteurs de données, certains producteurs ne passent pas directement par le CASD mais utilisent le boitier CASD pour la partie confidentialité. Donc, dans ces cas-là, il fallait attendre toutes les réponses de chaque producteur.
Et puis, tous les changements par rapport au projet initial étaient dus au coût. Ce n'est pas donné d'avoir tous ces accès et donc, j'ai toujours choisi le prix minimal sur l'accès, il n'y avait que moi, et la configuration minimale au niveau de l'ordinateur.

Et la configuration minimale au final, maintenant que tu l'utilises, est-ce qu'elle te convient ou tu as des regrets ?
Cela me convient parfaitement. C'est la configuration dont on n'a pas vraiment l'habitude, principalement au niveau de Office mais après, c'est pareil.

5'46 : Comment se présente son environnement de travail ?

Ton environnement personnel de travail, il se présente comment ?

Comme un environnement classique Windows, la dernière version de Windows.
Ensuite il y a deux dossiers, un dossier avec uniquement les données et les fichiers, et un dossier personnel pour sauvegarder des notes et tous les back-up.
Ensuite, il y a tous les logiciels demandés c'est-à-dire Open Office, tous les logiciels de statistiques spécialisés.
Il n'y a pas la connexion Internet, ce n'est pas possible de brancher une clé USB par exemple.
Il est interdit par exemple d'avoir une imprimante connectée à l'ordinateur et bien sûr il n'est pas possible de faire une capture d'écran.

Et ça te gêne au quotidien ou tu t'y es habitué ?

Non, pas du tout parce que je travaille avec deux écrans : l'écran avec le CASD et mon écran personnel, si j'ai besoin de quelque chose sur Internet, sur les données ou s'il y a quelque chose dans le code qui ne marche pas, je cherche dans l'autre ordinateur.

Oui, tu peux regarder des tutoriels à côté.

Exactement, c'est ça.

7'30 : Quelle est la procédure pour exporter les résultats ?

Pour exporter tes résultats, il y a une procédure spécifique c'est ça ?

Oui, après avoir trouvé le bon export, il faut envoyer le script et le résultat au CASD.

Et tu envoies ça comment ? Tu envoies un mél ?

Il y a une interface qui permet d'envoyer tous les exports directement à la « centrale » du CASD. Ce sont eux qui se chargent de valider.
Ensuite, quand on fait cette procédure, ils vont envoyer la décision par mél. A ce moment-là ils t'envoient la version Office ou une version du logiciel si tu travailles dans une version de logiciel spécifique par exemple LaTex, ils t'envoient la version en LaTex.

C'est la version qu'ils ont modifié ?

Exactement, c'est la même version que tu envoies, c'est la même chose que tu vas recevoir, la plupart du temps c'est Office. C'est la première chose qu'il faut demander.
Ensuite, ça dépend de la base de données ; par exemple, il y a des bases de données où c'est plus direct quand tu fais l'export : il y a une fenêtre qui s'ouvre et automatiquement, c'est validé ou pas. Dans les 5 minutes tu as directement les méls sans passer par la centrale du CASD.

Et toi tu as déjà eu des remarques sur tes scripts informatiques, sur tes sorties ?

Pas encore, je n'ai pas encore fait les sorties. J'attends d'être sûr parce qu'il y a une limite de sorties ou d'exports chaque année.

C'est ça, tu avais droit à 20 exports lors du premier abonnement dans la grille tarifaire, tu avais droit à un package ?

Oui, tout à fait.

Tu prends tes précautions pour ?

Oui, même si 20 est suffisamment large par rapport à ma thèse ou par rapport à ma recherche. Je préfère, on ne sait jamais, faire les bonnes choses et de ne pas dépasser les 20 exports.

Il faut être prudent.

9'49 : Dans la grille tarifaire, il est mentionné qu'un pack de 20 exports est inclus lors du premier abonnement. Qu'est-ce que contient un export ?

C'est quoi exactement un export ? Tu m'as dit ton script informatique plus une sortie ?

Exactement, ça fait partie de l'export mais c'est surtout la sortie (la table, un graphique) qui va compter parce que l'idée est qu'ils ne vont pas juger si tes résultats sont bons. Ce n'est pas leur travail. Ils regardent que l'on respecte toutes les règles de confidentialité, toutes les règles de base de l'enquête ou de la base de données.

C'est si tu respectes bien tes engagements.

Exactement, toutes les parties que tu avais signées à la base, il faut que cela soit représenté dans tes exports.

OK merci Deivyd.

10’33 : Qu’est-ce qu’une séance d’enrôlement ?

Avant d’accéder aux données confidentielles, il faut suivre une séance d’enrôlement ? C’est quoi exactement, ce n’est pas très rassurant ce nom ?

Cette séance, c’est pendant de 2 ou 3 heures une présentation du CASD : comment obtenir toutes les données, comment utiliser le boitier mais surtout, le plus important, c’est la partie confidentialité et secret statistique. Parce que finalement on a des données très nominatives soit au niveau territorial, soit par panel de secteurs d’activité.
Donc l’idée est apprendre comment utiliser les bonnes statistiques pour éviter tout repérage de chaque acteur.
Donc, cette séance est bien sûr obligatoire mais elle est très nécessaire pour la recherche.

C’est une formation aux bonnes pratiques pour utiliser les données auxquelles tu as accès.

Exactement.

Aujourd’hui, quel regard portes-tu sur la procédure d’accès ?

Pour moi, elle est évidente, je suis d’accord avec cette procédure, même si c’est très complexe au début. Il faut tenir compte qu’il y a toute la partie de confidentialité et aussi toute la partie du secret entre les producteurs et le CASD.

11’46 : L’importance de la gestion du temps

Y a-t-il une question que tu t’attendais à ce que je te pose et qui n’a pas été abordée ?

Surtout la partie de la durée, d’avoir le boitier.
Qu’est-ce qui se passe si on aimerait avoir plus de temps pour analyser les données parce que finalement on n’a accès uniquement pour 12 mois à partir du moment où l’on reçoit les données.

Et donc, ces 12 mois, il faut faire attention à bien faire l’installation du boitier. C’est ça qui prend un peu de temps et après, tu as 12 mois pour faire l’analyse.

Tout à fait.

12’24 : Le message de Deivyd aux futurs utilisateurs du CASD

Est-ce que tu as un message court à faire passer à de futurs utilisateurs ?

Il faut d'abord bien trouver la question et ne pas hésiter à aller demander directement au CASD parce qu'il y a beaucoup de sources, de bases de données qui s'actualisent ou qui arrivent presque tous les mois. Donc, il faut être très attentif.

Et est-ce que tu as l'impression de travailler sur des données riches qui te permettent de faire vraiment des recherches innovantes ?
Oui, tout à fait parce qu'avec ce degré de finesse, on peut élaborer des nouvelles recherches et surtout ajouter de nouvelles choses à toutes les recherches qui sont déjà faites.

13'11 : Remerciements et conclusion

Merci Deivyd.
Juste une petite question : il vient d'où ton joli accent ?

Je suis colombien.

Merci Deivyd, on a vu grâce à toi qu'il était intéressant, possible de travailler sur des données très fines au niveau des acteurs. Toi, c'est tout ce qui est entreprise qui t'intéresse au niveau de la région Nouvelle Aquitaine. Et puis on voit bien que la séance d'enrôlement, ce n'est pas si sorcier que ça.
Merci encore et bon courage pour la suite de ton travail.

Merci, au revoir.