Recueil du Symposium 2016 de Statistique Canada
Croissance de l'information statistique : défis et bénéfices

Application moderne de soumission des travaux pour accéder aux données administratives et aux données de recherche d'enquêtes confidentielles de l'IAB

Johanna Eberle, Dana Müller et Jörg Heining[1]

Résumé

L'Institute for Employment Research (IAB) est le service de recherche de l'Agence fédérale allemande de placement. Par l'entremise du Centre de données de recherche (FDZ) à l'IAB, des données administratives et des données d'enquête sur les personnes et les établissements sont fournies aux chercheurs. En collaboration avec l'Institute for the Study of Labor (IZA), le FDZ a mis en œuvre l'application de soumission des travaux (JoSuA), qui permet aux chercheurs de soumettre des travaux, en vue du traitement des données à distance grâce à une interface Web personnalisée. Par ailleurs, deux types de fichiers de sortie produits pour l'utilisateur peuvent être reconnus dans l'environnement JoSuA, ce qui permet de fournir des services d'examen de la divulgation plus rapides et plus efficaces.

Mots-clés : traitement des données à distance; microdonnées; confidentialité.

1. Données et modes d'accès au FDZ

Le Centre de données de recherche (FDZ) fournit des ensembles de microdonnées confidentiels de l'Institute for Employment Research (IAB) de façon uniformisée à des établissements de recherche (non commerciaux). On offre à la fois des ensembles de données administratives et de données d'enquête. Les données sont des ensembles de microdonnées de nature délicate concernant des personnes, des ménages et des établissements. Les données administratives proviennent d'avis d'emploi envoyés à des fournisseurs de services de sécurité sociale et de renseignements sur les prestations de chômage touchées, de la recherche d'emploi enregistrée et de la participation aux programmes du marché du travail, ainsi que des plans de formation, telles que recueillies par l'Agence fédérale allemande de placement. Les données recueillies dans le cadre de ces processus sont couplées au niveau individuel, de sorte que chaque individu puisse être suivi au fil du temps, tout au long de son historique d'emploi et de chômage. Toutes les données sont disponibles sur une base quotidienne, et l'année la plus ancienne de collecte est 1975. L'IAB mène aussi plusieurs enquêtes exhaustives sur différents sujets dans les domaines de la recherche d'emploi et les domaines connexes. Ces données sont considérées comme aussi confidentielles que les données administratives. En outre, le FDZ offre des données couplées employeur-employé, ainsi que des ensembles de données qui relient des données administratives et des données d'enquête (Heining, 2010).

L'accès aux données est uniformisé et dépend du degré d'anonymat de celles-ci, qui va des données absolument anonymes à des données très détaillées et peu anonymes[2]. En ce qui a trait au sujet du présent document, seul ce type

[1]Johanna Eberle, Institute for Employment Research, Regensburger Str. 100, D-90478 Nuremberg, johanna.eberle2@iab.de; Dana Müller, Institute for Employment Research, Regensburger Str. 100, D-90478 Nuremberg, dana.mueller@iab.de; Jörg Heining, Institute for Employment Research, Regensburger Str. 100, D-90478 Nuremberg, joerg.heining@iab.de

[2]Parmi les autres types de données disponibles au FDZ figurent les fichiers de campus et ceux que l'on appelle les fichiers à usage scientifique. Les fichiers de campus sont absolument anonymes et sont utiles pour l'enseignement, mais pas pour une analyse substantielle valide. Les fichiers de campus peuvent être téléchargés par les utilisateurs enregistrés qui acceptent les modalités d'utilisation. Par contre, les fichiers à usage scientifique sont constitués de données factuelles anonymes et sont préparés de façon particulière pour l'accès à distance. En dépit du degré élevé d'anonymat et de la réduction du niveau de détail des variables, ainsi que de la censure des valeurs extrêmes, les

de données est décrit de façon plus détaillée. Les ensembles de données peu anonymes ne comportent pas d'identificateur direct, comme le nom ou l'adresse, mais affichent un risque élevé de dés-anonymisation, en raison de la quantité et de la précision des caractéristiques visées. Par conséquent, l'accès aux ensembles de données peu anonymes est possible uniquement sur place ou par le traitement à distance.

Afin de pouvoir accéder aux données dans le cadre de l'utilisation sur place et du traitement des données à distance, certaines conditions doivent être respectées, conformément aux règlements (pour plus de détails, voir Hochfellner et coll., 2014). Le FDZ offre des formulaires de demande uniformisés pour les différents modes d'accès aux données, qui précisent si un projet de recherche se conforme à ces conditions. Une fois la demande de données acceptée, FDZ et l'établissement de recherche concluent une entente particulière pour l'utilisation des données. Les données peuvent être utilisées uniquement pour un projet particulier à l'intérieur d'une période déterminée énoncée dans le contrat.

Au FDZ, les ensembles de données peu anonymes sont conservés dans des serveurs de fichiers dans un réseau isolé. Pour l'utilisation sur place, les chercheurs invités ont accès à ce réseau par l'entremise de clients locaux qui n'ont pas accès à Internet ou à un autre réseau de l'IAB. Par ailleurs, FDZ fournit des locaux additionnels pour les chercheurs invités, à l'intérieur de cet environnement informatique sécurisé, à différents endroits en Allemagne et à l'étranger (Bender et Heining, 2011). Les chercheurs peuvent utiliser directement les données peu anonymes, mais ils ne peuvent pas télécharger ou transmettre les résultats eux-mêmes. Les résultats sont plutôt soumis aux chercheurs après examen de la divulgation par le personnel du FDZ (pour plus de détails sur la protection des données au FDZ, voir Hochfellner et coll., 2012).

Comparativement à l'utilisation sur place, le traitement des données à distance signifie que les chercheurs préparent leurs programmes au moyen de données d'essai artificielles et soumettent ces scripts au FDZ. Les programmes sont traités au moyen d'un logiciel statistique, et les résultats sont fournis aux chercheurs, après examen de la divulgation par le personnel du FDZ. Au cours du traitement des données à distance, les chercheurs n'ont pas accès directement aux données originales.

Avant 2015, le traitement des données à distance au FDZ n'était que partiellement automatisé. Les chercheurs envoyaient leurs scripts par courriel au FDZ, et le personnel du FDZ copiait manuellement ces scripts dans un réseau informatique isolé. Là, les scripts étaient traités avec les données. Une fois les exécutions de programmes terminées, le personnel du FDZ récupérait manuellement les résultats pour un examen de la divulgation. Toutefois, au cours des dernières années, le FDZ a connu une augmentation substantielle du nombre de travaux soumis pour le traitement à distance, soit un total d'environ 1 800 travaux soumis à distance en 2014. Par conséquent, la gestion du traitement des données à distance et la fourniture de services d'examen de la divulgation pour un nombre sans cesse croissant de travaux devenaient de plus en plus difficiles et longues, à l'intérieur de ce système partiellement automatisé. Afin de résoudre cette situation, FDZ a décidé de mettre en œuvre l'environnement JoSuA (application de soumission des travaux), qui sera décrit dans le paragraphe qui suit.

2. Présentation de JoSuA (application de soumission des travaux) au FDZ

L'ensemble logiciel JoSuA a été mis en œuvre au FDZ en avril 2015. Le logiciel a été élaboré et est maintenu par l'Institute for the Study of Labor (IZA). Afin de respecter les exigences particulières des données confidentielles de l'IAB et d'adapter le logiciel aux procédures organisationnelles du FDZ, certaines composantes de JoSuA ont été modifiées ou élargies au moyen de caractéristiques spéciales. À l'heure actuelle, tous les processus pertinents pendant le traitement des données à distance se déroulent dans l'environnement JoSuA, qui exécute la plupart des processus automatiquement. Seules certaines tâches, comme les services d'examen de la divulgation, nécessitent toujours l'intervention du personnel du FDZ.

fichiers à usage scientifique comprennent plus de renseignements que les fichiers de campus et peuvent par conséquent être utilisés pour une analyse empirique utile. Les fichiers à usage scientifique sont transmis aux établissements de recherche en vertu d'une entente sur l'utilisation des données.

Du point de vue de l'utilisateur, le traitement des données à distance est maintenant assuré exclusivement par l'entremise de l'interface Web JoSuA (https://josua.iab.de). Les chercheurs entrent dans l'interface Web et téléchargent leurs scripts. Mis à part un fureteur Web standard, le chercheur n'a besoin d'aucun logiciel supplémentaire. Les travaux qui ont été soumis au serveur Web sont par la suite transmis au serveur informatique. Ces serveurs sont dans le réseau sécuritaire et isolé du FDZ et donnent accès au serveur de fichiers dans lequel sont entreposées les données de nature délicate. Les travaux sont traités dans ce réseau sécuritaire au moyen du logiciel statistique Stata. Une fois l'exécution du programme terminée, les fichiers de sortie sont passés en revue pour le contrôle de tout risque de divulgation. Enfin, tous les fichiers de sortie approuvés sont mis à la disposition des chercheurs au moyen de l'interface Web.

En contrepoint de l'interface Web, les employés du FDZ utilisent une interface d'exploitation qui leur permet de mener diverses tâches administratives, comme la supervision de l'exécution des travaux, le lancement ou l'interruption des processus, ou encore la gestion des comptes d'utilisateurs et de projets. Cette interface fournit aussi des fonctions utiles pour l'examen de la divulgation.

En date de mai 2016, environ 200 projets (faisant habituellement intervenir plusieurs chercheurs) utilisaient activement la plateforme JoSuA pour le traitement des données à distance. La moyenne mensuelle de travaux soumis est d'environ 700. Comparativement à moins de 2 000 travaux par année, le nombre de travaux est maintenant environ quatre fois plus élevé que le nombre mensuel de travaux en 2014.

3. Produits internes et externes

Jusqu'à maintenant, la valeur ajoutée de cette nouvelle application de soumission des travaux se limite à une solution technique plus avancée. La principale innovation conçue pour JoSuA est la distinction entre les produits internes et externes, et les modes distincts de soumission des travaux qui sont offerts. Cela vient du fait que seulement une petite proportion des produits créée pendant le traitement des données à distance et fournie aux chercheurs est utilisée dans les faits dans une publication ou une présentation. La grande majorité des produits est créée à des fins internes de projet seulement. Par conséquent, une grande quantité de produits mise à la disposition des chercheurs est utilisée seulement pour vérifier les résultats de la préparation des données ou pour évaluer différents types d'analyses et de modèles.

Par suite de cette nouvelle distinction, les chercheurs peuvent maintenant faire un choix entre deux modes de soumission des travaux. Le premier mode, « Présentation/Publication », est conçu pour les résultats qui seront utilisés dans les faits dans une présentation ou une publication. Dans ce mode, les fichiers de sortie sont passés en revue manuellement pour le contrôle de tout risque de divulgation par les employés scientifiques du FDZ, et ils sont par la suite rendus disponibles comme fichiers de texte à télécharger par l'entremise de l'interface Web. Si les chercheurs ont l'intention de publier les résultats de leurs analyses des ensembles de données de l'IAB dans une présentation ou une publication spécialisée, un chapitre de livre, etc., ils soumettent leurs travaux au moyen de ce mode. Tout produit qui ne répond pas au critère d'anonymat absolu n'est pas diffusé (voir Hochfellner et coll., 2014).

Le deuxième mode, « Utilisation interne », peut être sélectionné pendant toutes les étapes préparatoires du processus de recherche décrit précédemment. Dans ce mode, les résultats peuvent seulement être pré-visualisés dans l'interface Web JoSuA. À cette fin, les fichiers de texte de sortie initiaux sont convertis en fichiers d'images, puis sont affichés dans le menu de visualisation. Il n'est pas possible de télécharger les fichiers d'images. Étant donné que les résultats sont disponibles uniquement dans JoSuA, le contrôle manuel de la divulgation est remplacé par un examen automatisé de la divulgation fondé sur le script. À cette fin, on utilise un script d'anonymisation des produits qui filtre les résultats selon des expressions courantes. Par ailleurs, les chercheurs sont autorisés à utiliser les résultats d'un travail « pour usage interne » uniquement pour préparer leurs analyses. Une fois un nouveau travail soumis en mode « Publication/Présentation », les résultats des travaux passés pour « Usage interne » ne sont plus accessibles. Étant donné que les scripts d'anonymisation n'englobent jamais tout, les risques de divulgation qui restent dans les résultats de consultation seulement sont couverts par une entente sur l'utilisation des données conclue entre l'établissement de recherche et le FDZ. Dans la fenêtre des résultats pour consultation seulement du mode « Utilisation interne », une mention en filigrane figure en arrière-plan pour rappeler aux utilisateurs ayant conclu l'engagement contractuel de ne

pas faire de saisie d'écran ou d'image des résultats et de ne pas mettre les résultats internes à la disposition d'un tiers, en dehors de l'entente sur l'utilisation des données du projet.

L'un des principaux avantages de cette distinction est que seuls les produits nécessaires pour une présentation ou une publication sont fournis aux chercheurs. Tous les autres produits générés dans le cadre du projet demeurent dans un environnement informatique sécurisé du FDZ. Cela augmente de toute évidence la sécurité des données et minimise aussi le risque de divulgation. Par ailleurs, étant donné que les travaux en mode « Utilisation interne » ne nécessitent pas de contrôle manuel de la divulgation, les résultats sont disponibles plus rapidement que dans le mode « Présentation/Publication ». Le fait de limiter le volume de produits à vérifier au chapitre des risques de divulgation comporte par conséquent un aspect important d'efficience. Dans le contexte d'un nombre croissant de projets de recherche, le fait de mettre l'accent sur le produit qui est effectivement publié garantit que le volume d'examen manuel de la divulgation est limité et peut être traité rapidement et efficacement.

Jusqu'à maintenant, environ 15 % des travaux sont soumis par les utilisateurs en mode « Présentation/Publication ». Les 85 % qui restent des travaux sont soumis en mode « Utilisation interne ». Le contrôle de la divulgation est assuré par le personnel scientifique et prend beaucoup de temps. En moyenne, en 2015, le contrôle de la divulgation a pris environ 16 minutes par tâche, dans une fourchette allant d'une minute à trois heures. Au premier trimestre de 2015, le temps total consacré au contrôle de la divulgation a été d'environ 120 heures. Après la mise en œuvre de la distinction entre les produits internes et externes, le premier trimestre de 2016 affiche maintenant un nombre total réduit de 90 heures.

4. Conclusion

Même si la mise en œuvre de JoSuA a signifié des changements et des ajustements pour toutes les parties concernées, elle peut être considérée comme un véritable succès. Les avantages, tant pour le FDZ que pour les chercheurs, de ce nouveau système pour le traitement à distance dépassent de loin le coût des ajustements.

L'un des principaux avantages de JoSuA est que les résultats du traitement des données à distance sont fournis plus rapidement aux chercheurs. Les résultats ne pouvant pas être téléchargés sont presque instantanément accessibles après la fin de l'exécution du programme. En outre, JoSuA permet le traitement à distance d'un nombre plus important de travaux. Comme mentionné précédemment, le nombre de travaux mensuels a quadruplé. En accélérant le processus de recherche et en éliminant les inefficiences et les frictions dans le processus, JoSuA comporte par conséquent des avantages fondamentaux pour la communauté scientifique.

Par ailleurs, étant donné que le nombre de produits transférés se limite maintenant aux 15 % des travaux qui ont été soumis en mode « Présentation/Publication », la charge de travail pour l'examen manuel de la divulgation est réduite. Si l'on tient compte du fait que, parallèlement, le nombre total de travaux a augmenté, le nombre de travaux à examiner manuellement est encore réduit d'environ 25 % par rapport au nombre précédant la mise en œuvre de JoSuA.

Bibliographie

Bender, S., and J. Heining (2011), "The Research-Data-Centre in Research-Data-Centre approach: A first step towards decentralised international data sharing", *FDZ-Methodenreport*, 07/2011 (en).

Heining, J. (2010), "The Research Data Centre of the German Federal Employment Agency: data supply and demand between 2004 and 2009", *Zeitschrift für ArbeitsmarktForschung*, 42, No. 4, pp. 337-350.

Hochfellner, D., Müller, D., Schmucker, A., and E. Roß (2012), "Data protection at the Research Data Centre", *FDZ-Methodenreport*, 06/2012 (en).

Hochfellner, D., Müller, D., and A. Schmucker (2014), "Privacy in confidential administrative micro data – implementing statistical disclosure control in a secure computing environment", *Journal of empirical research on human research ethics*, 9, No. 5, pp. 8-15.