Institut des Politiques Publiques

RAPPORT DE FIN DE TRAVAUX - Juillet 2024

Collecte des barèmes législatifs des politiques publiques

Mahdi Ben Jelloul
Paul Dutronc-Postel
Emmanuel Raviart

Institut des
Politiques Publiques

L'Institut des politiques publiques (IPP) a été créé par PSE et est développé dans le cadre d'un partenariat scientifique entre PSE-École d'Économie de Paris et le Groupe des écoles nationales d'économie et de statistique (GENES). L'IPP vise à promouvoir l'analyse et l'évaluation quantitatives des politiques publiques en s'appuyant sur les méthodes les plus récentes de la recherche en économie.

www.ipp.eu

Institut des
Politiques Publiques

RAPPORT DE FIN DE TRAVAUX - Juillet 2024

Collecte des barèmes législatifs des politiques publiques

Mahdi Ben Jelloul
Paul Dutronc-Postel
Emmanuel Raviart

ISSN 1951-8110

Les auteurs du rapport

Paul Dutronc-Postel Paul a rejoint l'IPP en 2018 après une thèse en économie du développement. Après avoir travaillé sur les thématiques liées à la fiscalité des ménages et à l'emploi et sur le développement du modèle de micro-simulation TAXIPP, il a fondé le programme Environnement de l'IPP.
Page personnelle : http://www.ipp.eu/annuaire/paul-dutronc/

Mahdi Ben Jelloul est économiste à l'IPP. Après avoir travaillé à France Stratégie, il intègre l'IPP en novembre 2014. Il participe au développement de l'outil de microsimulation OpenFisca et aux développements des modèles de microsimulation statique (TAXIPP) et dynamique (TAXIPP-Life) de l'IPP. Il réalise le module « dépendance » du modèle TAXIPP-LIFE.
Page personnelle : http://www.ipp.eu/annuaire/mahdi-benjelloul/

Emmanuel Raviart

Développeur Python et Javascript ayant travaillé à Etalab (data.gouv.fr, OpenFisca), à Jailbreak (DBnomics) et à l'Assemblée nationale (LexImpact). Il a déjà collaboré avec l'IPP sur des projets couvrant OpenFisca et les barèmes IPP.

Remerciements

Ce projet n'aurait pas pu être mené à bien sans le travail méticuleux et l'investissement remarquable de Pauline Bouvet tout au long de son stage de master à l'Institut des politiques publiques d'avril à juillet 2023.

Les auteurs tiennent à remercier les nombreux contributeurs aux barèmes législatifs des politiques publiques issus des barèmes IPP et des paramètres présents dans OpenFisca, qu'ils viennent de l'IPP ou de la communauté OpenFisca, notamment de la cellule LexImpact de l'Assemblée nationale pour sa coopération et la MSA pour ses multiples contributions à la mise à jour des prestations sociales. Merci aussi aux contributeurs de la Drees et de la DSS pour leurs divers apports concernant les retraites .

SOMMAIRE

Introduction

Contexte du projet

Le projet CoBaLPP - Collecte des barèmes législatifs des politiques publiques a été retenu dans le cadre du troisième appel à projets collaboratifs lancé par le GIS CollEx-Persée le 15 juin 2021 et clos le 10 décembre 2021. Il est porté par PSE - Ecole d'économie de Paris et co-piloté par Colette Cadiou (DipSO - INRAE) et Mahdi Ben Jelloul (Institut des politiques publiques, PSE - Ecole d'économie de Paris), responsables scientifiques au nom du coordinateur administratif.

Rappel des objectifs du projet

Ce projet de collecte cherche à répondre à un besoin des chercheurs et de certaines administrations publiques de disposer d'une source de référence facilement mobilisable. Pour ce faire, il vise à établir de façon pérenne une base de données des barèmes législatifs des politiques publiques, exhaustive, documentée et annotée, facilement réutilisable, collaborative et ouverte, en s'appuyant sur une première collecte initiée par l'Institut des politiques publiques. Le projet a ainsi deux dimensions principales, l'une sur le fonds de données récoltées et abondant la base de données, et l'autre sur les outils techniques permettant la pérennisation, la réutilisation et l'accessibilité de la base de données.

Ainsi, d'une part, un effort important de saisie a été entrepris pour améliorer la

couverture temporelle des données et améliorer les métadonnées, notamment compléter et fiabiliser les références législatives. D'autre part, des structures permettant la maintenance, la contribution collaborative, et la réutilisation des données ont été mis en place, afin de favoriser la diffusion, l'accessibilité et la fiabilité des données. Le développement d'un outil de validation permet d'établir un diagnostic clair de l'existant pour mieux le compléter. L'introduction d'une interface dédiée a permis d'améliorer la quantité et la qualité des contributions actuelles par les contributeurs naturels de la base de données, mais aussi de mobiliser des experts hors de la communauté initiale des contributeurs.

Enfin, afin de faciliter la découvrabilité des données, les séries de paramètres ont été indexées et le corpus a été rendu interrogeable à l'aide de son intégration dans un portail de données ouvertes (DBnomics) qui dispose de capacités de visualisation et d'une interface de programmation (API) permettant le téléchargement des séries par les logiciels usuels de statistiques. En d'autres termes, ceci permet aux chercheurs, statisticiens et autres utilisateurs de la base de données d'accéder à celle-ci directement lors de l'exécution de leurs programmes informatiques et de l'intégrer facilement à leur travail. Enfin, la base de données comme les outils développés sont tous diffusés sous une licence libre ouverte et sont librement réutilisables.

Ce rapport présente le travail et le principe de la complétion des données législatives réalisé dans le cadre du projet (Chapitre 1). Il rappelle ensuite succintement les outils et projets utilisant la base de données (Chapitre 2). Il revient sur la structure et le contenu des données, et sur les outils et fonctionnalités développés dans le cadre du projet (Chapitre 3). Enfin, il présente les efforts de diffusion et d'amélioration de l'accessibilité des données (Chapitre 4), ainsi que les étapes engagées pour la construction d'une communauté d'utilisateurs-contributeurs experts (Chapitre 5).

Enfin, un plan de gestion des données est annexé à la fin de ce document.

CHAPITRE 1

COMPLÉTION DES DONNÉES LÉGISLATIVES

Un effort conséquent a été fourni pour actualiser les données des divers pans de la législation ayant encore cours, tant en termes de valeurs que de références.

1.1 Enrichissement des méta-données des données existantes

Un travail a été mené pour compléter les méta-données des données existantes, notamment en fournissant des références législatives quand elles manquaient ou quand elles étaient incomplètes, soit par croisement avec une autre source de données existantes (les paramètres législatifs d'OpenFisca, un calculateur sociofiscal) soit par ajout de nouvelles références.

Le travail de complétion a bénéficié des contributions de la cellule LexImpact de l'Assemblée nationale. En effet, cette équipe utilise le logiciel libre de microsimulation OpenFisca afin de fournir aux députés une interface web dénommée LexImpact [1] permettant de réaliser des simulations paramétriques. Elle est donc tout par-

1. https://leximpact.an.fr/

ticulièrement intéressée par les références législatives reliées aux paramètres modifiables. L'unicité du format des paramètres ainsi que la volonté de mutualisation des efforts de complétion et de convergence vers une source unique a permis une collaboration fluide, car grandement facilitée par le recours aux outils développés dans le cadre de Collex-Persée (validateur, documentation).

Les méta données déjà existantes (références législatives, date de parution au journal officiel, documentation), ont ainsi été complétées et homogénéisées dans leur format.

De plus, une métadonnée supplémentaire a été ajoutée, à savoir la date de dernière validité confirmée d'un paramètre. Celle-ci, renseignée lors de la validation d'une modification de la base de données, permet d'en assurer la qualité à chaque réutilisateur.

1.2 Données complétées

1.2.1 Dans le passé

Un effort substantiel a été conduit pour essayer de compléter les références législatives manquantes en partant des plus récentes et en remontant dans le temps pour de très nombreux paramètres dont notamment ceux ayant encore cours et vérifier systématiquement la validité des URL associées.

En particulier, les champs des prélèvements obligatoires et des prestations sociales sont complétée.

À la date du 1er juillet 2024, la base de données représente ainsi 10 Mbit de données (hors historique des versions), pour environ 5 000 paramètres, 25 000 valeurs uniques, et 11 000 références législatives ; ceci peut être comparé aux 3 700 paramètres, 18 500 valeurs uniques, et 2 200 références législatives de la base de données à la date du 1er janvier 2021. Un considérable travail de complétion et

d'assainissement a donc été accompli.

L'ensemble des cotisations sociales, impôts et prestations déjà présentes ainsi que les paramètres intervenant dans le calcul des retraites ont été complétées jusqu'à leurs plus récentes modifications.

1.2.2 Nouvelles données

De nouveaux paramètres ont été également ajoutés. Ces ajouts sont de différents types. Premièrement, certaines séries de paramètres étaient déjà présents, parfois de façon incomplète (références manquantes, etc) dans OpenFisca et d'autres sont issues des paramètres récoltés par les chercheurs de l'IPP pour différents travaux sans être présents dans la base des barèmes IPP. Deuxièmement, de nouveaux pans de la législation comme les droits de succession, les impôts pesant sur les sociétés et la taxe d'habitation, de nombreux autres paramètres ayant trait aux impôts et aux prestations sociales, notamment des aides au logement et des aides destinées aux jeunes ont été intégrés dans la base de données.

Ils sont listés ci-dessous en respectant l'arborescence de la base et leur provenance est indiquée entre parenthèses.

- droits-de-mutation (nouveau)
- taxation_societes (nouveau)
- taxe_habitation (nouveau)
- impot_revenu
    - calcul_impot_revenu
        - plaf_qf
            - quotient_familial (parts supplémentaires, OpenFisca)
- prestations_sociales

- invalidite
  - pensions_invalidite (nouveau)
- aides_logement
  - action_logement (OpenFisca)
    - locapass
    - mon_job_mon_logement
    - visale
  - logement_social (OpenFisca)
    - plu
    - plai
- education (OpenFisca)
  - aide_permis_pro_btp
  - alimentation
  - carte_des_metiers
  - contrat_engagement_jeune
  - depart1825
  - garantie_jeunes
  - garantie_pret_etudiant
  - gratuite_musees_monuments
  - mobilite
    - internationale
    - master
    - mobili_jeune
    - parcoursup

- pass_culture
- sante_psy

- minima_sociaux (OpenFisca)
  - accident_travail (OpenFisca)

1.2.3 Dans le futur

L'objectif du projet Collex Persée étant de garantir la pérennité et la mise à jour constante de la base de données, un effort important a été consacré à la préparation du futur de la base, et des technologies pouvant être mises en oeuvre pour l'assurer.

Ainsi, la date de dernière preuve de validité a été mise à jour et permettra de détecter plus facilement si une valeur n'est pas à jour dans le futur. Grâce à l'inclusion de ce nouveau paramètre en effet, le travail d'harmonisation et d'homogénéisation des bases de paramètres législatifs des barèmes IPP avec ceux en provenance d'OpenFisca à permis d'exploiter pleinement une innovation de l'équipe LexImpact de l'Assemblée nationale. En effet, la cellule LexImpact a mis au point un outil d'automatisation de vérification des paramètres en vigueur permettant de mettre à jour la date de validité de la dernière valeur en cours la plus récemment vérifiée, ou de modifier le paramètre si la valeur a changé et que la nouvelle valeur est retrouvée. L'outil développé a été utilisé pour suggérer des mises à jour [2] à travers des demandes de fusion soumise sur le dépôt des paramètres qui doivent néanmoins toujours être validés par des humains. Un aperçu de ces travaux a été présenté [3] à la communauté OpenFisca et les détails du travail conduit sont aussi disponibles [4].

2. https://git.leximpact.dev/leximpact/exploration/fiscal-qa/-/blob/master/dataset_generation/liste_des_PR_OF-FR_ouvertes.md?ref_type=heads
3. https://cloud.leximpact.dev/index.php/s/J7RSLGeZQM2eZ9X
4. https://git.leximpact.dev/leximpact/exploration/fiscal-qa/-/blob/master/dataset_generation/Readme.md?ref_type=heads

CHAPITRE 2

TRAVAUX UTILISANT LES BARÈMES LÉGISLATIFS DES POLITIQUES PUBLIQUES

De nombreux travaux ont recours aux barèmes législatifs des politiques publics, notamment des calculateurs et des simulateurs socio-fiscaux. Nous en présentons les plus emblématiques.

2.1 Travaux conduits à l'IPP

Les barèmes législatifs des politiques publics, diffusés par l'IPP sous le nom de "barèmes IPP" sont une composante essentielle du calculateur OpenFisca-France qui équipe le modèle de microsimulation de l'IPP, TAXIPP [1]. Ce modèle de microsimulation a été utilisée pour de nombreuses études conduites à l'IPP [2] dont pour les conférences annuelles du budget où l'IPP présente ses travaux sur l'impact de la loi de finances sur les ménages à des membres de la commission des finances de l'Assemblée nationale qui les discutent en public.

Le calculateur des pensions de retraite françaises développé par l'IPP en utilisant l'architecture OpenFisca, OpenFisca-France-Pension puisent également ses pa-

1. https://www.ipp.eu/methodes/taxipp-outils/
2. https://www.ipp.eu/liste-de-lensemble-des-publications-utilisant-taxipp/

ramètres législatifs dans les barèmes législatifs des politiques publics.

2.2 Travaux conduits hors de l'IPP

Plusieurs outils de simulation économique sont développés en dehors des institutions portant le développement de la base de données. Ceci témoigne, d'une part, du vif intérêt porté à la base de données par des utilisateurs experts. D'autre part, cela illustre l'importance des outils techniques développés dans le cadre du projet, dont l'intention est avant tout prospective, et de garantir la robustesse des processus de contribution. Enfin, cela contribue à solidifier une communauté diverse et active de contributeurs, qui peuvent alimenter la base de données et contribuer à sa pérennité. Les deux principaux outils faisant appel à la base de données sont présentés succintement ci-dessous.

2.2.1 Simulateur LexImpact de l'Assemblée nationale

La cellule LexImpact de l'Assemblée nationale utilise également le calculateur OpenFisca-France pour réaliser des simulations socio-fiscales[3] à destination des députés, mais également ouvertes au grand public avec certaines restrictions (nécessité de vérifier que les résultats sont conformes au secret statistique). A ce titre, elle est l'un des principaux utilisateurs et contributeurs à la base de données.

2.2.2 Modèle trajectoire de la Drees

Le modèle Trajectoire, élaboré par la Drees (Direction de la recherche, des études, de l'évaluation et des statistiques, une direction de l'administration centrale des ministères sanitaires et sociaux), est un modèle de microsimulation dynamique dont l'objectif principal est de produire des projections des populations futures de

3. https://socio-fiscal.leximpact.an.fr/?budget=true&parameters=irpp_economique

retraités et de leur niveau de pension [4]. Les dernières versions du modèle utilisent, pour le calcul des pensions de retraite, les paramètres législatifs des barèmes législatifs des politiques publiques, diffusés par l'IPP.

4. https://www.insee.fr/fr/statistiques/1305197?sommaire=1305205

CHAPITRE 3

STRUCTURATION, OUTILLAGE ET MAINTENANCE

3.1 Structure des données

Les données sont hébergées au format YAML sur le dépôt de partage public de l'IPP sur la forge GitLab. Le format YAML a été retenu, car il permet de représenter de façon lisible des informations élaborées comme une combinaison de listes et de dictionnaires imbriqués. Un schéma bien précis a été élaboré avec des champs obligatoires et d'autres facultatifs. Tous ces champs doivent également suivre une structure plus ou moins stricte.

Les paramètres choisis doivent représenter la loi le plus fidèlement possible. Chaque évolution d'une valeur doit être justifiée par une référence législative. Ainsi, afin de respecter ce principe tout en réduisant au strict minimum la charge pour le contributeur, seuls trois champs sont obligatoires lors de la création d'un paramètre :

- la description longue du paramètre
- les valeurs du paramètre
- la référence législative

Les champs dits facultatifs rassemblés dans les autres métadonnées sont les suivants :

- La description courte
- La description longue en anglais
- La date de validité de la dernière valeur en cours la plus récemment vérifiée
- La date de publication au journal officiel
- L'unité
- La documentation, de portée générale
- Les notes, annexes apportant des précisions sur une valeur à une date donnée notamment

Il est important de noter que ce format est compatible avec celui des paramètres consommés par le logiciel de microsimulation OpenFisca. La version de ce dernier qui modélise le système socio-fiscal français peut donc bénéficier directement des mises à jour de la base de données. Ce format est par ailleurs utilisé par les dérivations d'autres pays modélisés avec OpenFisca qui peuvent donc bénéficier tant de la structuration du format que des outils de validation.

3.2 Outils

3.2.1 Validation

Les modifications ou ajouts à la base de donnée atterrissent sur le dépôt sous forme d'une nouvelle branche soumise pour intégration.

Chaque branche est scrutée par un validateur [1] qui détecte les éventuelles erreurs de format, vérifie l'existence des liens, etc, et les notifie (figure 3.1). Le validateur permet de localiser précisément l'erreur tout en précisant autant que faire se

1. https://control-center.tax-benefit.org/

peut la source de l'erreur, notamment en quoi la modification apportée diffère du format attendu. Le contributeur peut alors très rapidement effectuer un diagnostic et apporter son correctif. Bien entendu, la validation finale, dès lors qu'elle passe les tests automatiques, n'est effectuée que par un humain dûment mandaté.

3.2.2 Contribution

Afin de favoriser la contribution d'experts qui ne serait pas familier avec les outils de développement informatique et les forges logicielles en particulier, une interface de contribution a été réalisée. Elle permet à un contributeur, régulier ou occasionnel, d'amener ses modifications ou ajouts via une interface conviviale. Cette interface permet d'effecteur une recherche dans l'ensemble des paramètres existants, d'identifier et de sélectionner le paramètre à modifier. L'ensemble des données et métadonnées est alors modifiable. Les champs décrivant le paramètre tant en français qu'en anglais sont le premier par ordre d'apparition afin de ne laisser aucun doute au contributeur (figure 3.2).

Viennent ensuite les valeurs datées dont sont modifiables la valeur elle-même, mais également son unité, ainsi que les références législatives à la source du changement de valeur et les notes pouvant éclairer les conditions de cette évolution (figure 3.3)

Par ailleurs, cette interface permet de soumettre des modifications qui sont *de facto* au bon format en mutualisant l'infrastructure logicielle du validateur.

Enfin, le contributeur est invité à s'identifier. S'il le désire, il peut laisser son nom ou son adresse mail afin de se signaler comme source fiable, mais surtout se signaler au validateur final. S'il dispose déjà d'un compte GitHub ou GitLab, il peut s'authentifier en un clic (figure 3.4).

Un dernier module dépliable permet d'avoir un aperçu du fichier YAML modifié (voir au bas de la figure 3.4).

FIGURE 3.1 – Centre de contrôle des contributions (branches soumises à validation)

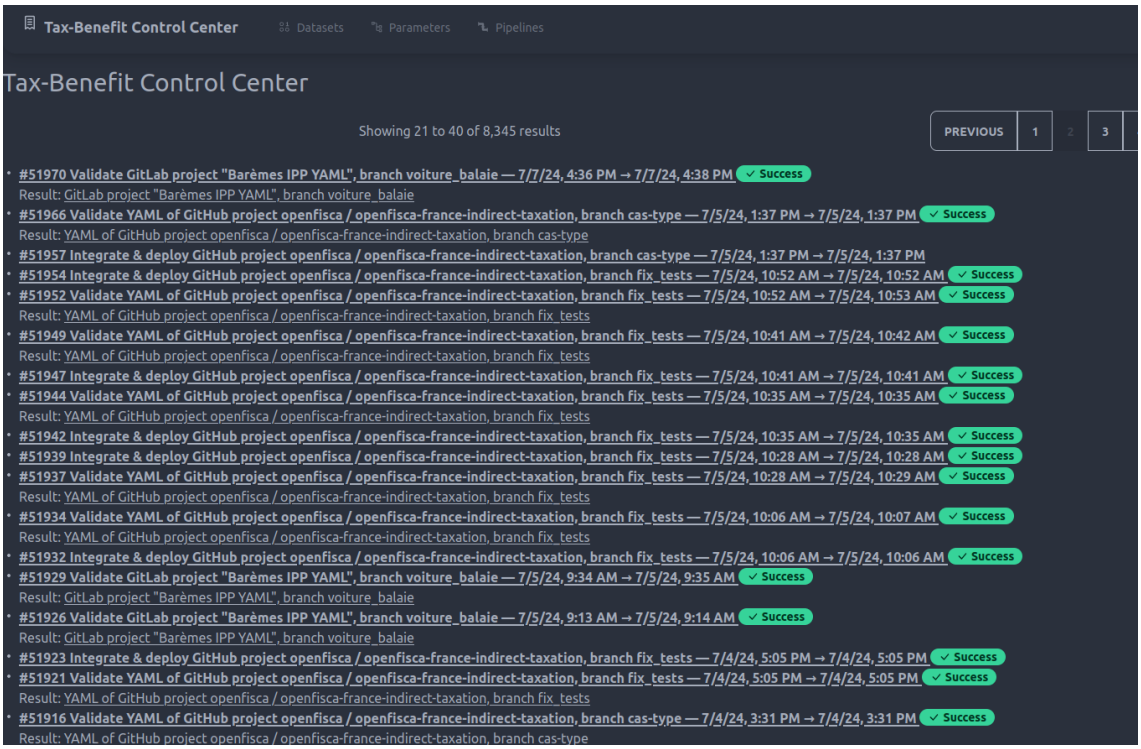


FIGURE 3.2 – Édition de la description

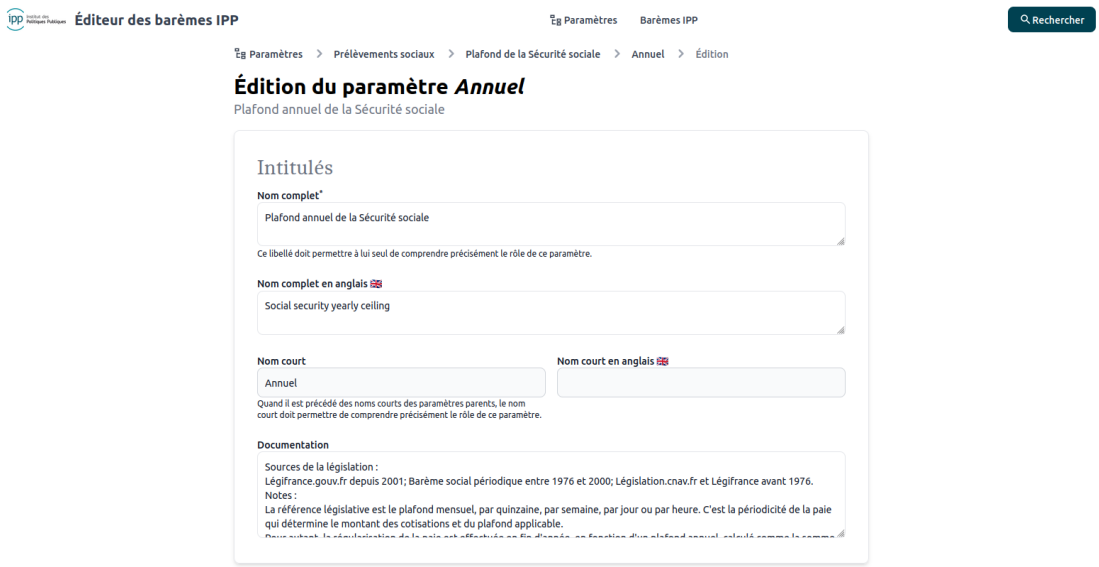
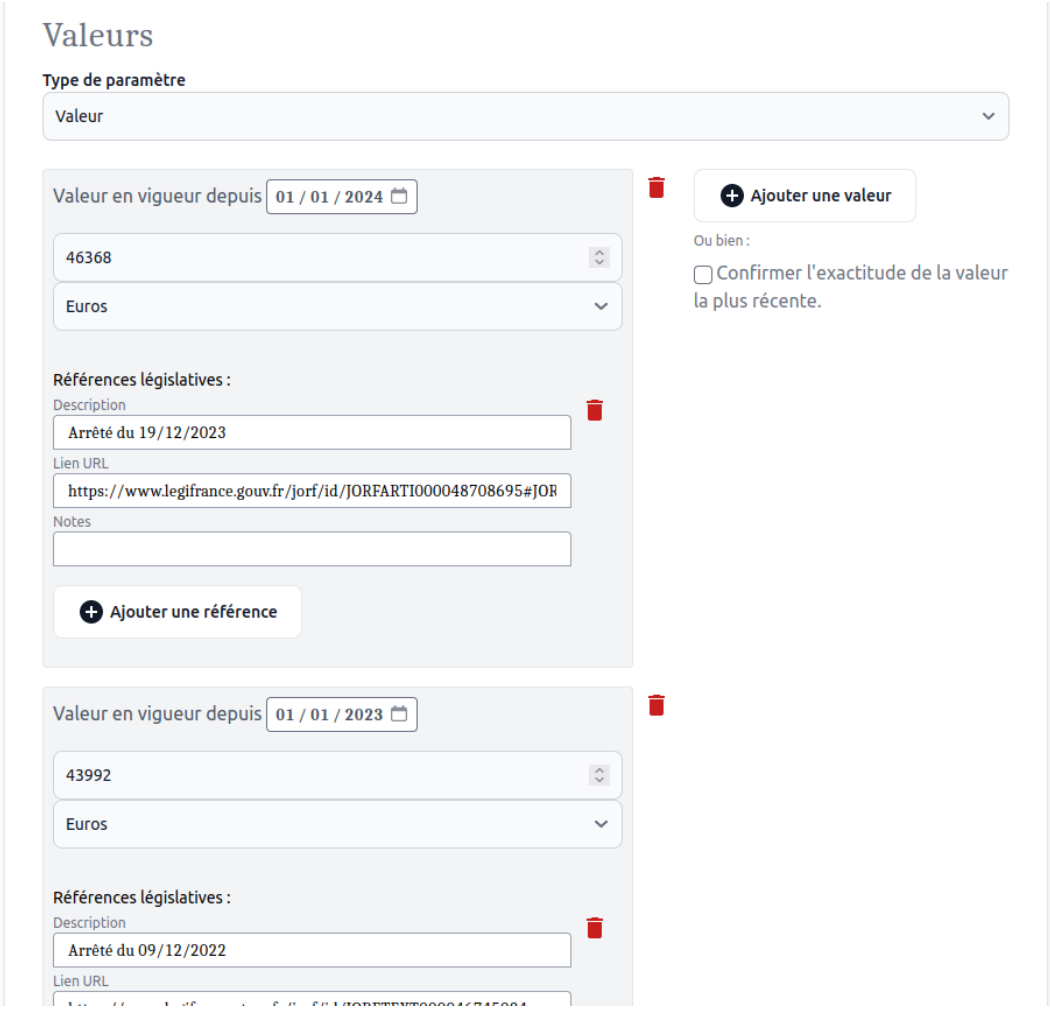


FIGURE 3.3 – Édition des valeurs

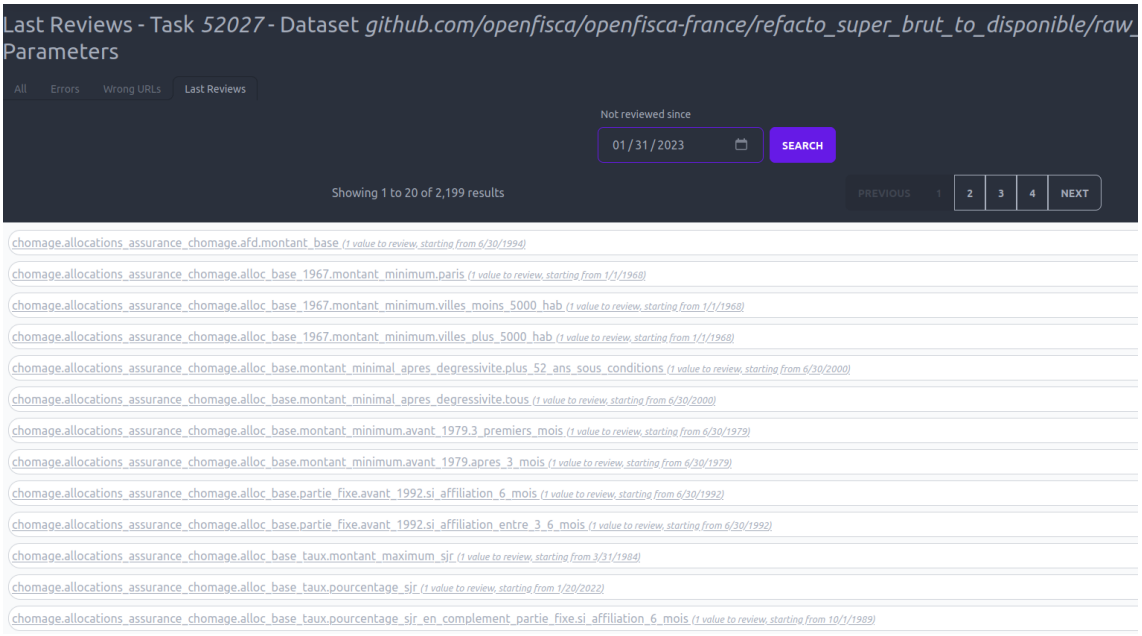


3.2.3 Tableaux de bord

Des modules, préfigurant un véritable tableau de bord qui va au-delà d'un simple validateur, ont également été développés. Ils permettent de repérer certaines anomalies, par exemple, détecter les URL invalides, mais aussi repérer les paramètres qui n'ont pas été vérifiés récemment alors qu'ils ont toujours cours (figure 3.5).

FIGURE 3.4 – Authentification du contributeur

FIGURE 3.5 – Authentification du contributeur



CHAPITRE 4

DIFFUSION

4.1 Diffusion

4.1.1 Nouveau site web Barèmes IPP

Le site des barèmes IPP [1] accueillent les tableaux qui sont produits à partir de la base de données organisés par thèmes. Les modifications ont été intégrées au fur et à mesure sans que le site web soit arrêté. La consultation des tableaux permet d'accéder aux paramètres sous divers formats et l'utilisateur peut aussi être redirigé vers la page de DBnomics correspondantes. S'il constate une erreur, il peut également, directement depuis le tableau, accéder à l'interface de contribution pour proposer une modification.

4.1.2 DBnomics

L'ensemble des paramètres des barèmes législatifs des politiques publiques sont désormais disponibles sur DBnomics qui fait office d'agrégateur de séries et de plateforme de diffusion. Ses données sont mises à jour quotidiennement et chaque révision est archivée. Son site web permet de recherche des séries temporelles par

1. https://www.ipp.eu/baremes-ipp

mot-clé ou par dimension (figure 4.1). Elle permet évidemment de télécharger les séries, notamment au format CSV ou XLSX. Elle permet aussi d'avoir aperçu de la série sous forme de tableau ou de graphique (figure 4.2). Les séries qui sont indexées de manière unique sont facilement mobilisables via les API disponibles pour les logiciels usuels de statistiques comme R, Stata ou SAS, et les langages de programmation comme Python, Julia ou Matlab. Un exemple d'utilisation de l'API Python de DBnomics est visible à la figure 4.3.

FIGURE 4.1 – Recherche dans l'interface DBnomics

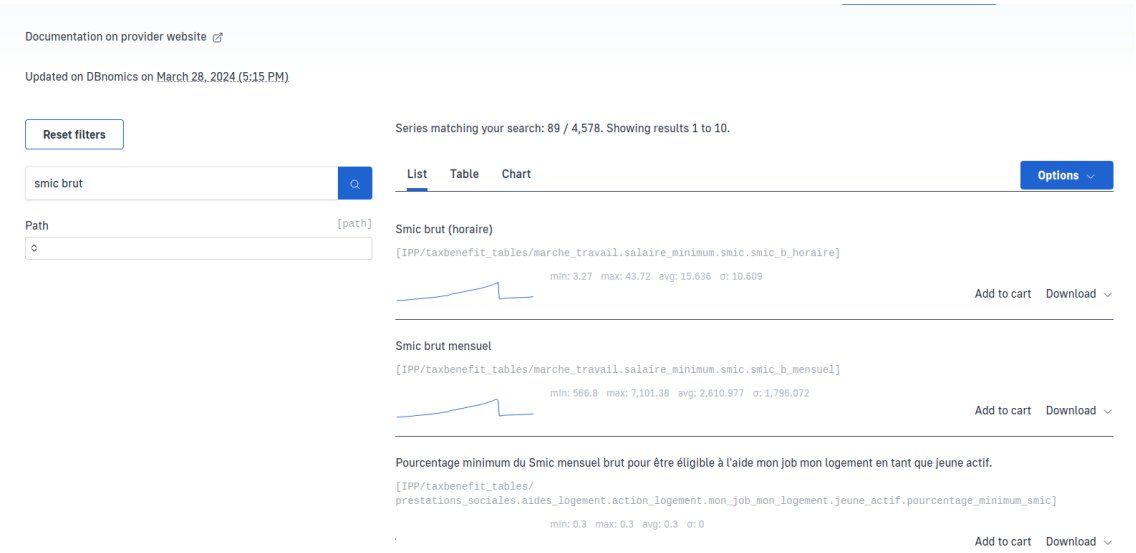


FIGURE 4.2 – Aperçu de la série du Smic brut

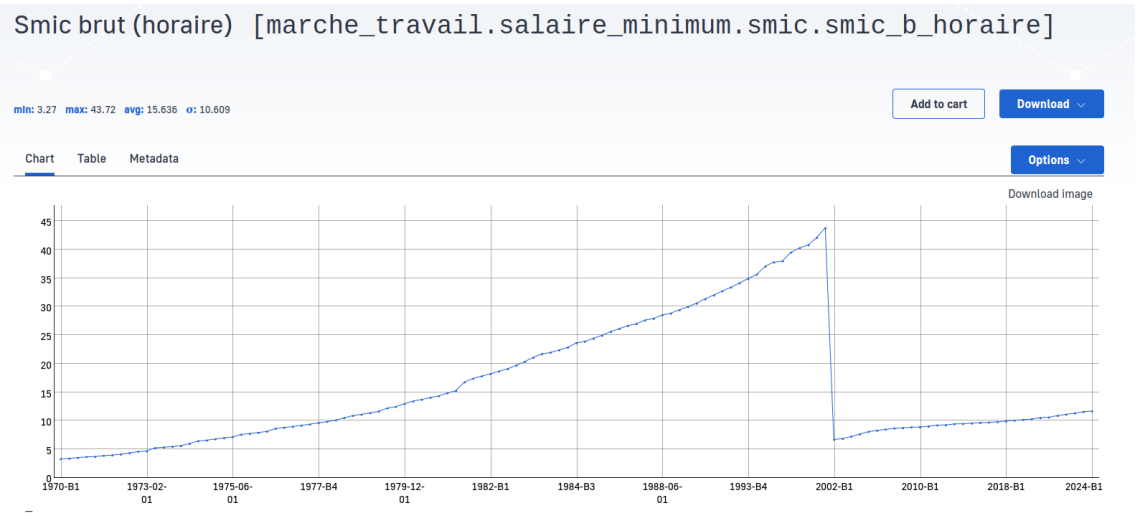


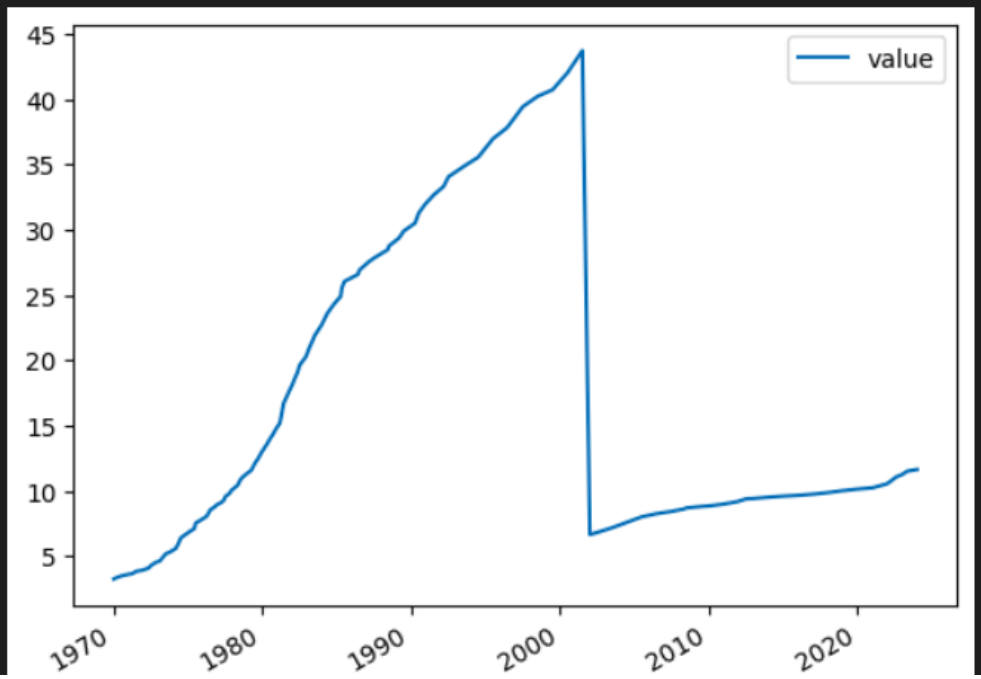
FIGURE 4.3 – Exemple d'utilisation de l'API Python de DBnomics

```
##
import dbnomics


df = dbnomics.fetch_series(
    api_base_url = 'https://api.ipp.staging.db.nomics.world/',
    provider_code = 'IPP',
    dataset_code = 'taxbenefit_tables',
    series_code = 'marche_travail.salaire_minimum.smic.smic_b_horaire',
    )


df[["period", "value"]].set_index("period").plot()
```

```
<Axes: xlabel='period'>
```



CHAPITRE 5

COMMUNICATION

Les travaux entrepris dans le cadre du projet Collex-Persée ont donné lieu à une communication constante sur leurs objectifs et l'état d'avancement, notamment auprès contributeurs les plus actifs et ceux qui présentent un fort potentiel de contribution. Elle a également été tournée vers les utilisateurs actuels et ceux en devenir, notamment via DBnomics. Il est utile de rappeler que, souvent, ces deux types d'acteurs, contributeur et utilisateur, peuvent se confondre en un acteur profitant de l'ensemble des fonctionnalités offertes par la plateforme tant en amont qu'en aval.

5.1 Lancement du projet

Un effort de prise de contact et de sensibilisation avait été pris avec les acteurs les plus importants avant le lancement du projet proprement dit. Ils ont ainsi pu suivre le déroulement du projet sans difficultés de compréhension. L'ensemble du travail étant conduit sur la forge GitLab des barèmes IPP mais aussi sur les forges GitHub des projets proches, l'information sur les développements conduits et leurs finalités a pu circuler de manière fluide et les retours de toutes les parties prenantes pris en compte.

5.2 Une attention particulière pour les contributeurs

Le lien avec les contributeurs, anciens comme nouveaux, a été consolidé par une pratique inclusive. Les échanges qui ont eu lieu sur les forges ont permis des retours sur les contributions et la fourniture de précisions sur le remplissage des champs et même des échanges sur la localisation des références législatives. La rédaction d'une documentation détaillant la contribution a été conduite par itération en prêtant une attention particulière aux questions émanant des contributeurs.

5.3 Constituer une communauté

Une communication des objectifs et des outils, une fois ceux-ci bien avancés, plus englobante, a aussi été conduite afin que l'ensemble des acteurs puissent se rencontrer et fasse mûrir ou émerger des problématiques plus transverses. Un événement rassemblant près de 30 personnes de la Drees, de la DSS, de la CAF, de l'Assemblée nationale sans oublier l'IPP a été organisé en ce sens à l'Ecole d'économie de Paris le 8 mars 2023. Il a permis de présenter l'éventail des réalisations attendues.

Il a été suivi d'un atelier technique avancé mêlant les membres de l'équipe IPP avec ceux l'équipe LexImpact de l'Assemblée nationale pour finaliser des travaux d'homogénéisation de la base de données nécessitant des efforts de clarification et de coordination substantiels.

L'embryon initial constitué des chercheurs de l'IPP et d'autres intervenants autour d'OpenFisca a débouché sur une coopération accrue entre l'IPP et la cellule LexImpact. Notre espoir est qu'un tel niveau de coopération puisse être atteint à terme avec les autres contributeurs, notamment les organismes publics. Un évènement prévu à la rentrée de septembre 2024 sera organisé en ce sens.

Annexe A : Plan de gestion des données

A.1 Description des données, collecte et réutilisation de données existantes

A.1.1 Recueil et production des données originales, réutilisation de données préexistantes

Les données consistent en des "paramètres" législatifs, c'est-à-dire des valeurs numériques qui sont utilisées dans les textes de loi et réglementaires pour mettre en œuvre les formules des prélèvements obligatoires et des transferts sociaux. Un bon exemple est celui du barème de l'impôt sur le revenu, qui est constitué des seuils des différentes tranches d'imposition marginale et des taux d'imposition correspondants. La chronologie de l'impôt sur le revenu peut ainsi être appréciée à travers celle de son barème (et par le suivi des autres nombreux paramètres qui contribuent à la formule de calcul de l'impôt).

Les données sont recueillies et produites sous forme de fichiers YAML, et stockées dans un répertoire de données en ligne, hébergées sur la forge Gitlab de l'IPP [1].

Ainsi, toutes les données sont préexistantes au projet, et sont pour une majo-

1. (https://gitlab.com/ipp/partage-public-ipp/baremes-ipp

rité d'entre elles, publiées au Journal Officiel de la République française (JORF). D'autres sont recueillies dans d'autres types de documents officiels, tels que des circulaires ministérielles, mais lorsque c'est possible, le JORF est préféré comme source. Un principe directeur de collecte de la base de données est précisément de ne contenir que des paramètres et données qui sont présents tels quels dans la loi. On s'abstient ainsi de collecter des valeurs "transformées" ; par exemple, pour un paramètre qui est calculé au prorata de la base mensuelle des allocations familiales, on collectera la valeur du prorata spécifiée dans la loi, par opposition à la valeur en euros obtenue par le calcul.

La provenance des données est documentée par des métadonnées précises et exhaustives. Ainsi à chaque valeur sont associées une, ou (lorsque nécessaire), plusieurs références législatives, constituées d'un intitulé (précisant le cas échéant le ou les articles pertinents), un lien hypertexte permanent vers le texte de loi numérique, et, de façon facultative, une date de publication au JORF.

Le logiciel de gestion de version Git et la forge GitLab sont utilisés pour permettre :

- le suivi explicite des modifications de la base de données au cours du temps
- les arbitrages à rendre sur les modifications de la base de données et les discussions entre le responsable scientifique et autres contributeurs disposant d'autorisations spécifiques sur la base de données, et les différents contributeurs.

Plusieurs outils sont par ailleurs développés pour permettre la gestion des données, et se rapportent à :

- le formatage et l'assainissement de la base de données
- la visualisation de la base de données
- l'édition de la base de données
- la réutilisation de la base de données par des individus ou des tiers

- le suivi dynamique de la base de données

Ces outils sont décrits dans la section suivante.

A.1.2 Nature des données produites

Les données sont des valeurs numériques, auxquelles sont associées des métadonnées. Elles sont stockées dans une bibliothèque numérique selon des critères et règles précises de classement.

L'unité de base de stockage des données est un fichier YAML, qui contient l'ensemble des valeurs prises par un paramètre donné au cours du temps. Ainsi, de la base mensuelle des allocations familiales, du taux global de contribution sociale généralisée sur les pensions de retraites, du salaire minimum interprofessionnel de croissance, ou de l'abattement conjoint pour les donations entre vifs. Un changement, même mineur, de définition du concept légal, entraîne la création d'un paramètre spécifique, stocké dans un nouveau fichier ; par exemple, des taux encadrant le calcul de la valeur de la nue propriété en fonction de l'âge de l'usufruitier, pour lesquels les catégories d'âge, ayant changé de définition en 2004, entraînent la création de nouveaux paramètres (de "moins de 21 ans révolus" à "moins de 20 ans révolus", de "moins de 31 ans révolus" à "moins de 30 ans révolus", etc).

Un paramètre est ainsi défini par un concept inchangé utilisé par la législation, dont la valeur est susceptible d'être modifiée au cours du temps (notamment à l'occasion des lois de finance et de financement de la Sécurité sociale). Dans la base de données, à chaque valeur d'un paramètre donné est associée une date, qui constitue sa clef d'identification au sein du fichier recensant toutes les valeurs du paramètre. Cette date correspond à la date d'application de la valeur en question (par opposition à sa date d'entrée en vigueur, ou la date à laquelle elle a été votée ou décidée). Au sein du fichier, les valeurs sont stockées par ordre chronologique de leur clef.

Les fichiers des différents paramètres sont stockés dans une arborescence de dossiers qui permet de retrouver facilement les différentes valeurs. Cette arborescence classifie les fichiers en grandes "sections", "sous-sections", etc.

Les métadonnées sont indexées ou bien au niveau de chaque valeur (et référencées par la date-clef d'identification de celle-ci), ou bien au niveau du nœud de l'arborescence pertinent (par exemple, la section du Code général des impôts se rapportant à un ensemble de paramètres). Elles sont stockées au sein du fichier correspondant (fichier d'un paramètre pour les métadonnées se rapportant à une valeur ou à un paramètre, fichier d'un nœud pour les métadonnées se rapportant à un nœud).

L'utilisation du format YAML retenu se justifie par ses caractéristiques techniques, et par les usages prévus pour la base de données. En effet, le format YAML est un format qui associe :

- la légèreté (économie d'espace de stockage)
- la lisibilité
- un usage très répandu et la compatibilité avec de nombreux langages de programmation (de façon à pouvoir permettre la consommation de la base de données par de nombreuses applications)

Il combine ainsi deux atouts majeurs vis-à-vis des publics visés et de la communauté des contributeurs à la base de données :

- il est largement utilisé dans les communautés qui alimentent et consomment la base de données au premier chef (développement de microsimulateurs et calculateurs socio-fiscaux, au premier rang desquels OpenFisca)
- il est très facile à convertir sous d'autres formats, comme la visualisation sous forme de tableaux en ligne [2] ou sous forme de fichier au format CSV.

2. www.ipp.eu/baremes-ipp/

La base de données est évolutive et participative et a donc vocation à continuer à croître au-delà de la fin du projet Collex-Persée. À la date du 1er juillet 2024, elle représente 10 Mbit de données (hors historique des versions), pour environ 2 500 paramètres, et plus de 25 000 valeurs uniques.

Enfin, plusieurs programmes informatiques complètent la base de données en la dotant d'outils visant à la rendre plus robuste, plus facilement améliorable, et plus facilement découvrable et utilisable.

Le "validateur" de la base de données participe de la robustesse technique et de la qualité formelle de la base de données. Il identifie et corrige les erreurs de format pour les données et les métadonnées. Il intervient principalement, une fois la première version de la base de données établie, en amont des contributions qui visent à modifier ou mettre à jour la base de données (cf. section A.2.2).

Le "tableau de bord" explore la base de données et rend compte de sa qualité. Il permet ainsi d'identifier

- les fichiers et valeurs de la base de données pour lesquelles des métadonnées facultatives sont absentes
- les fichiers et valeurs de la base de données dont la qualité doit être vérifiée. Celles-ci sont identifiées sur la base de la date de dernière vérification de validité d'une valeur théoriquement toujours en vigueur.
- les fichiers et valeurs de la base de données dont le rythme de mise à jour normal suggère qu'une mise à jour doit être effectuée. Ainsi, peu après la date anniversaire de modification d'un paramètre, la validité de la dernière valeur renseignée pour celui-ci peut être contrôlée.

Le "visualisateur" est un outil informatique qui consomme les données stockées dans la forge Gitlab de l'IPP et permet leur visualisation sur le site Internet de l'IPP, sous forme de tableur. Chaque série de données présente dans les tableurs visualisés comprend un bouton "Edition" qui ouvre une interface de contribution (cf. infra) et

permet de proposer des modifications ou des mises à jour des séries.

L' "interface de contribution" de la base de données est une interface utilisateur dédiée à l'édition des séries de valeurs. Elle s'adresse aux contributeurs et contributrices d'experts qui ne seraient pas familiers des outils de développement informatique et des forges logicielles en particulier. Elle permet à un contributeur, régulier ou occasionnel, d'amener ses modifications ou ajouts via une interface conviviale. Cette interface permet d'effecteur une recherche dans l'ensemble des paramètres existants, d'identifier et de sélectionner le paramètre à modifier. L'ensemble des données et métadonnées est alors modifiable. Une fois que l'ensemble des modifications a été effectué, la contribution peut être soumise à la validation de l'équipe en charge de la base de données (cf. infra ; section A.2.2). Alternativement à l'interface de contribution dédiée, les contributions à la base de données peuvent être effectuées par une modification, "experte", directe de la base de données au sein de la forge Gitlab de l'IPP. Les deux types de contributions sont soumises au même processus de validation, l'interface n'agissant que comme un intermédiaire technique.

La réutilisation de la base de données par des individus (contributeurs ou non), et en particulier par des applications tierces, est favorisée à travers son inclusion dans l'agrégateur DBNomics. DBNomics est un agrégateur et une plateforme de diffusion de séries temporelles de données économiques qui est très largement utilisé dans la communauté de recherche internationale en économie ; il s'agit d'un projet porté par le CEPREMAP, et soutenu par l'AFD, la Banque de France, France Stratégie et le réseau PROGEDO. L'inclusion de la base de données CoBalPP dans l'agrégateur DBNomics opère à plusieurs niveaux. Elle améliore d'une part la découvrabilité des données et leur capacité de diffusion. D'autre part, elle permet aux données de disposer des outils développés par l'équipe DBNomics pour la réutilisation des données accessibles via l'agrégateur : API web, paquets Python, R, Stata, SAS. Ainsi, les chercheurs souhaitant se référer à la base de données développée dans le cadre du projet CoBalPP peuvent utiliser les données directement dans leur

programme informatique et dans leur logiciel de statistique, en récupérant la série voulue en une ligne de commande.

A.2 Documentation et qualité des données

A.2.1 Nature des métadonnées et documentation

Les métadonnées collectées pour chaque valeur sont les suivantes.

- label : ce champ obligatoire est une description en français du paramètre. Il n'est pas limité en termes de caractères et doit être auto-suffisant, c'est-à-dire que cette description doit permettre de le distinguer le paramètre de façon unique dans le système socio-fiscal. Cela passe par le fait de rappeler l'ensemble des éléments de contexte du paramètre, ou a minima d'indiquer le dispositif dans lequel le paramètre intervient.
- label_en (optionnel) : le champ 'label_en' est affiché sur la version anglaise du site des Barèmes IPP. Il s'agit de la traduction en anglais du champ 'label'.
- short_label (optionnel) : la métadonnée 'short_label' répond au besoin de désigner un nœud ou un paramètre de la façon la plus signifiante et la plus courte possible. C'est donc un nom non ambigu avec un nombre limité de caractères. Ce nom court autorise les abréviations les plus connues. Il s'inscrit toujours dans un contexte de présentation reprenant l'arborescence du yaml, permettant ainsi de différencier deux paramètres ayant le même 'short_label'. Il peut servir dans des interfaces graphiques (sites, tableaux, graphiques, etc.)
- ipp_csv_id (optionnel) : Le champ 'ipp_csv_id' sert à la rétrocompatibilité de certains outils de l'IPP. Il a vocation à disparaître.
- last_value_still_valid_on (optionnel) : Ce champ permet de témoigner que la dernière valeur ('value') d'un paramètre est bien toujours en vigueur à la date 'YYYY/MM/DD' (Date de la relecture). Comme indiqué par son nom, cette

métadonnée ne concerne pas toutes les ‘value‘. Elle atteste uniquement que la dernière valeur a été vérifiée comme étant toujours en vigueur. Ainsi, la date de ce champ est forcément postérieure ou égale à la dernière valeur du paramètre (et ceci peut facilement être vérifié automatiquement). En pratique, il est impératif de fournir la référence législative à date affichant directement la dernière valeur à jour. Si cette ‘href‘ est une information nouvelle (par exemple, changement d'emplacement de l'article dans le code), le contributeur peut décider de l'ajouter dans les références au niveau de la dernière date de ‘value‘. On mettra l'‘href‘ à date uniquement dans la merge request dans le cas où il est déjà dans les références, afin d'éviter les redondances. Ce champ est optionnel et permet ainsi la capitalisation progressive de cette information dans la base de paramètres.

- unit (optionnel) : cette métadonnée indique l'unité du paramètre. Ce champ est optionnel et permet ainsi la capitalisation progressive de cette information dans la base de paramètres. L'ajout d'une unité permet souvent de raccourcir ou expliciter les champs ‘label‘ et ‘short_label‘.
- documentation (optionnel) : ce champ de texte libre contient des informations relatives au paramètre, notamment des explications générales, l'extension des acronymes, ou encore des références législatives plus larges que des valeurs spécifiques (par exemple Service-Public ou l'Urssaf).
- documentation_start (optionnel) : ce champ est booléen. Il vise à signaler le nœud le plus haut auquel on doit remonter pour construire une documentation pertinente en agrégeant les champs ‘documentation‘ intermédiaires.
- reference : Une référence législative ou réglementaire est exigée à chaque ajout ou modification d'une valeur, afin d'identifier la loi d'où provient le paramètre et ainsi permettre la revue des modifications implémentées. La règle est de fournir une référence avec laquelle l'on peut identifier la valeur concer-

née en un clic. Autrement dit, la valeur doit être écrite dans la référence législative transmise.

Ainsi, cette référence peut être :

— un lien direct vers l'article de loi en vigueur à date ;

— un lien vers le décret qui modifie la valeur (utiliser la version initiale du décret qui contient la valeur) ;

— ou, idéalement, les deux : l'article en vigueur et le décret à l'origine de la modification.

Il est impératif que la référence soit officielle (lien LégiFrance vers un article ou un décret de préférence, circulaire officielle, etc.) et il est préférable de sélectionner un article de loi codifié.

— title : un intitulé

— href : une URL

- official_journal_date (optionnel) : il se compose des dates de publication au Journal Officiel, pour chaque 'value'. S'il y a plusieurs publications par date, on fait un tableau '[2019-01-01, 2019-01-04]'.
- notes (optionnel) : le champ 'notes' contient des informations spécifiques à une date. Il est donc spécifique aux paramètres.

Comme indiqué dans la liste ci-dessus, deux métadonnées sont obligatoires pour qu'un point de données soit inclus dans la base de données : une référence législative complète, et une description succincte du paramètre (c'est-à-dire de la série de valeurs). Bien que facultatives, les autres métadonnées peuvent être analysées par des robots qui permettent d'identifier les zones de la base de données pour lesquelles les métadonnées sont incomplètes et donc de participer à contrôler et améliorer l'état de la base de données.

L'objet de la base de données étant inédit, il n'existe pas à proprement parler de standard de métadonnées établi, ce qui d'ailleurs ajoute à la contribution scienti-

fique du projet. Le format et la définition des métadonnées fait l'objet de discussions régulières entre les contributeurs à la base de données internes comme externes au projet. Les principes directeurs retenus pour les métadonnées sont : l'unicité de l'information et la concision, la précision, l'exhaustivité, la communauté d'usage, et la vérifiabilité.

La base de données comprend un guide détaillé de contribution à la base de données, à destination des contributeurs internes et externes. Elle comprend également une documentation des métadonnées, qui inclut des recommandations détaillées sur leur fonction et la manière de les renseigner, ainsi que des exemples.

L'arborescence de la base de données fait l'objet de discussions et de décisions collaboratives impliquant des contributeurs à différents échelons de responsabilité. Le premier échelon de l'arborescence est la responsabilité du responsable scientifique du projet. La responsabilité des sous-échelons de l'arborescence peut être déléguée à des responsables de "sections" ou "sous-sections" de la base de données. La modification de l'arborescence de la base de données intervient rarement et toujours de façon coordonnée avec les principaux réutilisateurs de la base de données, pour lesquels cette modification peut avoir des conséquences importantes (adaptation des programmes informatiques consommant la base de données, etc). Ainsi, plus l'échelon de l'arborescence concerné est élevé, plus les modifications doivent être rares, plus le nombre de contributeurs ayant accès au processus décisionnel est restreint.

A.2.2 Mesures et principes de contrôle de la qualité des données

Les mesures et principes de contrôle de la qualité des données sont de deux natures : de forme et de fond.

Le "validateur" est un outil informatique automatisé qui standardise le format

des données contenues dans la base de données à chaque instant. Il a été développé pour assainir la base de données pendant sa collecte, mais aussi de façon à garantir la qualité future de celle-ci. Il opère ainsi en amont de chaque ajout et modification à la base de données en vérifiant la compatibilité formelle de la modification avec le format préexistant d'une série de données. Notamment, les types de données et la présence des métadonnées obligatoires est vérifiée. Le résultat des tests formels lancés par le validateur conditionnent la possibilité de l'ajout des données : une barrière technique est ainsi opposée à la modification de toute donnée qui ne satisfait pas aux tests (absence de référencement, syntaxe incorrecte, etc). Concrètement, ce "validateur" est un exécutable dont l'exécution est déclenchée de façon automatique par la soumission de toute proposition de modification. De plus, le "validateur" permet, le cas échéant, de localiser précisément l'erreur tout en précisant autant que faire se peut la source de l'erreur, notamment en quoi la modification apportée diffère du format attendu. Le contributeur auteur de l'ajout proposé est alors en mesure d'effectuer un diagnostic et d'apporter un correctif à l'ajout proposé. Si les tests requis par l'exécution du "validateur" sont validés, la modification proposée est soumise à l'étape de validation sur le fond.

La validation sur le fond s'opère par un système de validation par les pairs. Chaque modification est soumise à l'approbation d'une relecture et une validation à travers l'interface Gitlab par un ou une membre-contributeur qui dispose des autorisations nécessaires. Cette étape de relecture et validation permet un échange ouvert et publiquement accessible au sujet de chaque décision d'ajout et de mise à jour. Une barrière technique s'oppose ainsi à toute modification non formellement relue et acceptée par un ou une membre-contributrice de la base de données qui dispose des autorisations nécessaires. Une structure organisationnelle est proposée à la communauté des contributeurs et contributrices qui organise les responsabilités et compétences des uns et des autres en matière de relecture et validation. Le responsable scientifique de la base de données propose notamment une délégation

des compétences de validation à certains responsables de "sections" de l'arborescence de la base de données, en fonction de leurs contributions passées et de leurs compétences.

Un ajout ou une mise à jour doit pouvoir être "vérifiable" facilement par son ou ses validateurs. Autant que faire se peut, la vérification d'une valeur doit pouvoir se faire "en un clic", ce qui fait écho au critère des métadonnées obligatoires (cf. section A.2.1) : une référence législative précise doit obligatoirement être présente pour chaque valeur, présence qui est vérifiée par le "validateur" en amont de la validation sur le fond.

A.3 Stockage et sauvegarde pendant le processus de recherche

A.3.1 Stockage et sauvegarde des données et métadonnées au long du processus de recherche

À l'issue du projet CoBalPP, les données sont stockées sur la forge Gitlab de l'IPP. Ceci assure la pérennité de la base de données, dans la mesure où les données stockées par l'IPP sur sa forge sont sauvegardées par Gitlab (l'IPP bénéficiant d'un abonnement "universitaire").

De plus, la structure décentralisée du logiciel de gestion de version Git garantit qu'une version complète et à jour des barèmes est stockée par chaque contributeur "expert" actif, ce qui participe grandement à la résilience de la base de données pour le cas (improbable) d'une mise en défaut complète du serveur Gitlab.

A.3.2 Sécurité des données et protection des données sensibles

La base de données ne contient pas de données sensibles. Elle est en accès libre, publiée sous licence libre de copie, de diffusion et d'utilisation.

Afin de limiter les contributions inadaptées ou frauduleuses et les attaques, toute contribution à la base de données doit être associée à un identifiant constitué d'un nom (qui peut être un pseudonyme) et d'une adresse email. Toute contribution initiale d'un contributeur ou d'une contributrice est soumise à l'approbation du responsable scientifique de la base de données ou d'un contributeur ou d'une contributrice qui dispose des autorisations nécessaires.

L'utilisation de la technologie de versionnement Git garantit de plus la traçabilité de toutes les modifications successives de la base de données, l'identification des auteurs et autrices des contributions, et la possibilité de pouvoir revenir facilement à tout état antérieur de la base de données.

A.4 Exigences légales et éthiques, codes de conduite

A.4.1 Données à caractère personnel

Aucune donnée à caractère personnel n'est collectée dans le cadre de ce projet par défaut.

Seule la contribution active à la base de données peut engendrer une collecte de données pouvant être considérées comme à caractère personnel. La base de données étant conçue comme collaborative, la contribution active à celle-ci fait partie intégrante du projet. C'est pourquoi la collecte de données personnelles dans le cadre d'une contribution active à la base de données est décrite ci-après.

Un identifiant constitué d'un nom (qui peut être un pseudonyme) et d'une adresse email est associé à chaque contribution, et est ainsi accessible aux gestion-

naires de la base de données. Cette adjonction s'opère ou bien par l'intermédiaire de l'utilisation d'un compte Gitlab ou Github (il s'agit alors de la pratique courante et commune à tous les projets collaboratifs hébergés par ce type de plateforme), ou bien en renseignant manuellement un nom (qui peut être un pseudonyme) et une adresse email de contact dans la plage dédiée de l'éditeur de paramètres. Cet identifiant permet à chaque contribution d'être discutée avec son auteur ou son autrice, et validée, tout en réduisant au maximum la quantité d'information nécessaire à cet échange.

Lors de la saisie d'une contribution, le contributeur qui choisit de renseigner son identifiant manuellement (et donc de ne pas utiliser un compte Github ou Gitlab) est informé du fait que son identifiant (pseudonyme et adresse email associée) sera stocké et associé à sa contribution, de façon à pouvoir tracer son auteur et échanger avec lui si nécessaire. Le fait de ne pas donner son accord pour un tel stockage bloque toute contribution.

A.4.2 Autres questions juridiques, titularité et droits de propriété intellectuelle sur les données

L'Institut des politiques publiques est le producteur et détenteur de la propriété intellectuelle sur la base de données produites par le projet Collex-Persée ; c'est-à-dire que les partenaires fondateurs de l'Institut des politiques publiques, la fondation PSE-École d'économie de Paris et le GENES, en sont copropriétaires. La base de données étant toutefois largement collaborative, la propriété intellectuelle de chaque contribution appartient à son contributeur, en l'absence de transfert explicite de la propriété intellectuelle.

La base de données est mise à disposition gratuitement, sous licence libre de copie, de diffusion et d'utilisation sous réserve de citation. Ceci s'étend à toutes les contributions et modifications faites à la base de données.

A.4.3 Questions éthiques et codes déontologiques

La base de données est ouverte et collaborative. Les contributions extérieures à l'équipe scientifique sont soumises à validation de forme et de fond (cf. section A.2.2), mais elles sont encouragées, les erreurs étant discutées et corrigées sous l'hypothèse que les contributeurs agissent de bonne foi. Symétriquement, les contributeurs de la base de données y contribuent de façon à en améliorer la qualité, en respectant les principes directeurs de contribution et les critères de qualité, et en acceptant que leurs contributions peuvent faire l'objet de modifications ou de refus si elles ne correspondent pas auxdits principes. Les échanges entre contributeurs, entre contributeurs et relecteurs des contributions, et entre les membres de la communauté en général, se font dans le respect des personnes, l'inclusivité et la collaboration.

La version de la base de données publiée et visualisée sur le site Internet de l'IPP correspond à un ensemble de valeurs de paramètres qui a effectivement ou a eu dans le passé une valeur légale. Il n'est, par exemple, pas fait état dans cette version publiée de valeurs discutées dans des projets de loi. Cependant, la base de données se présentant sous la forme d'un répertoire versionné, elle peut facilement être utilisée pour créer des versions alternatives, "prospectives", incluant des valeurs futures ou possibles des paramètres, qui peuvent par exemple être utilisées dans des simulations par des calculateurs socio-fiscaux.

A.5 Partage des données et conservation à long terme

A.5.1 Partage des données (méthode, temporalité)

La base de données est gratuite et ouverte, libre de copie, de diffusion et d'utilisation sous réserve de citation.

Elle est accessible à tout moment dans sa version à jour sous différents formats :

- téléchargement de la base de données intégrale sur le serveur ouvert de l'IPP (www.gitlab.com/ipp/partage-public-ipp/baremes-ipp-yaml)
- consultation sur le site Internet de l'IPP (www.ipp.eu/baremes-ipp/)
- téléchargement "par tables" sur le site Internet de l'IPP au format CSV
- acccès via le paquet Python et l'API DBNomics

Une fois passé le processus de validation sur la forme et le fond (cf. supra, section A.2.2), une modification ou mise à jour de la base de données est quasi-immédiatement déployée, et les différents formats sous lesquels la base de données est disponible sont également mis à jour automatiquement.

Au niveau de chaque paramètre (donc, de chaque fichier), la métadonnée 'last_value_still_valid_on' renseigne une "date de dernière validité", c'est-à-dire la dernière date à laquelle la valeur la plus récente du paramètre (la valeur dont la date de mise en œuvre est la plus récente) a été vérifiée et validée comme étant toujours en vigueur.

Les données sont facilement accessibles et découvrables :

- via le site Internet de l'IPP
- via la forge Gitlab de l'IPP
- via le paquet Python DBNomics

La base de données est maintenue, améliorée et mise à jour par les équipes de l'IPP en premier lieu ; et par les contributeurs et contributrices extérieures éventuelles. Sa nature gratuite, ouverte, libre de copie, de diffusion et d'utilisation sous réserve de citation est inaltérable.

A.5.2 Détermination des données à conserver, préservation à long terme

La base de données est un projet central au sein de l'IPP. Son accessibilité, sa mise à jour, l'amélioration de sa qualité et sa préservation à long terme font partie intégrante du projet de l'Institut, et constituent donc une priorité stratégique. La préservation à long terme de la base de données en son état d'accessibilité à l'issue du projet est donc l'objectif. La base de données est entreposée, a minima, au sein de la forge Gitlab de l'IPP, qui est un outil central de travail pour l'IPP, dont l'existence est liée à celle de l'IPP.

Le transfert de la base de données vers un autre entrepôt peut être envisagé (par exemple pour l'inclure dans un environnement de travail plus vaste, comme la forge Github d'OpenFisca) ; le cas échéant, l'IPP conserverait une copie "miroir" de la base de données dans sa propre forge (c'est-à-dire une copie exacte et immédiatement mise à jour). Pour le cas où l'IPP ne serait plus en mesure de garantir la pérennité de l'accessibilité de la base de données (c'est-à-dire, pour le cas où l'existence même de l'IPP serait remise en cause), une version à jour de la base de données sera entreposée dans le répertoire scientifique ouvert Zenodo (outre les versions déjà disponibles à cette date).

Par ailleurs, la base de données a vocation d'être abondée au fur et à mesure du temps, dans trois dimensions.

D'une part, malgré sa couverture déjà très large du système socio-fiscal français, certains dispositifs, notamment des dispositifs historiques disparus, peuvent encore ne pas avoir été recensés. Le cas échéant, une dimension d'amélioration concerne donc l'inclusion de ces paramètres existants encore non recensés sous forme de nouvelles séries dans la base de données. Autant que faire se peut, un tel ajout inclut toutes les valeurs depuis la création du dispositif.

D'autre part, l'amélioration de la base de données peut se faire par l'extension

des séries temporelles des paramètres, vers le passé comme vers le futur. Une extension vers le passé consiste à découvrir et abonder de nouvelles valeurs plus anciennes d'un paramètre déjà recensé. Ce cas de figure relève d'une incomplétude et est donc exceptionnel, étant donné l'objectif et les résultats du présent projet. Une extension vers le futur consiste à enregistrer de nouvelles valeurs pour les paramètres recensés.

A.5.3 Méthodes et outils logiciels nécessaires pour accéder et utiliser les données

Les formats sous lesquels les données sont disponibles sont multiples, ce qui garantit qu'un utilisateur potentiel dispose d'une manière facile d'y accéder et correspondant à son usage.

Ainsi, les données sont :

- consultables directement sur l'interface de visualisation sur le site Internet de l'IPP
- téléchargeables table par table au format CSV à partir de l'interface de visualisation sur le site Internet de l'IPP
- téléchargeables en intégralité sur la forge Gitlab de l'IPP, sous forme de dossier de fichier YAML
- 'clonables' à partir de la forge Gitlab de l'IPP
- consommables à travers l'utilisation du paquet Python ou de l'API DBNomics

Le minimum requis pour une première consultation de la base de données est donc un navigateur web (téléchargement ou consultation sur le site IPP), ou un autre type d'accès à Internet (autres accès détaillés ci-dessus). Si les données ont été téléchargées, un logiciel permettant de lire avec les formats CSV (LibreOffice, Microsoft Office, etc) et/ou YAML (VS Code, vim, etc) est nécessaire. Dans tous les

cas, un simple éditeur de texte suffit à accéder aux données.

A.5.4 Attribution d'un identifiant unique et pérenne

Un DOI est attribué à la base de données dans son ensemble, et sert pour la citation de la base données.

Le cas échéant, une clef d'indentification correspondant à chaque série est adjoint à ce DOI générique pour l'identification plus précise des séries.

A.6 Responsabilités et ressources en matière de gestion des données

A.6.1 Identité, rôle, position et institution de rattachement du responsable de la gestion des données

La responsabilité scientifique du projet est assumée par Mahdi Ben Jelloul, économiste à l'IPP, pendant la durée du projet, qui est portée par Mme Colette Cadiou (INRAE). Mahdi Ben Jelloul ayant quitté l'IPP à l'issue du projet, la responsabilité scientifique en est transférée à Paul Dutronc-Postel, économiste à l'IPP également. Le plan de gestion des données est élaboré conjointement par Mahdi Ben Jelloul et Paul Dutronc-Postel, avec l'aide du reste de l'équipe impliquée dans le projet, en particulier Emmanuel Raviart.

À l'issue du projet Collex-Persée, l'Institut des politiques publiques continue à mettre à jour, approfondir et améliorer la base de données, et à en garantir l'accessibilité dans le futur. La gestion et la responsabilité scientifique de la base de données, et notamment le respect et les évolutions éventuelles du présent plan de gestion des données, sont assurés par Paul Dutronc-Postel, ou par un ou une économiste IPP qui prend sa suite.

L'IPP agit, à l'issue du projet Collex-Persée, en coordonnateur de la communauté des contributeurs et réutilisateurs de la base de données, communauté dont la constitution était un des objectifs du projet. L'IPP prend la responsabilité du caractère réutilisable de la base de données à tout moment, et notamment

- de sa complétude
- de son exactitude
- de sa mise à jour de façon réactive
- de son accessibilité

Cet engagement s'applique aux données visualisables sur le site Internet IPP sous le nom de "Barèmes IPP".

L'IPP décide des dispositions les plus appropriées relatives au stockage et à la mise à disposition physique de la base de données auprès du public. Ceci peut signifier de transférer la base de données vers un environnement de travail plus vaste, comme la forge Github d'OpenFisca (cf. section A.5.1), étant bien entendu que (cf. section A.5.1) la nature gratuite, ouverte, libre de copie, de diffusion et d'utilisation sous réserve de citation de la base de données est inaltérable, et que (cf. section A.5.2) une version de la base de données est toujours disponible et conservée sur la forge Gitlab de l'IPP.

Le responsable scientifique de la base des données propose à des contributeurs, que ceux-ci soient ou non membres de l'équipe permanente de l'IPP, d'endosser des responsabilités dans la gestion de la base de données, et donc de se voir attribuer des autorisations correspondantes quant à la possibilité de modification de la base de données (droits "d'édition" et de "validation"). Ces propositions sont le fruit de discussions et d'échanges collégiaux au sein de la communauté des contributeurs et utilisateurs, que l'IPP participe activement à animer et faire vivre.

A.6.2 Ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable)

La découvrabilité, l'accessibilité, l'interopérabilité et la réutilisabilité sont au cœur du projet CoBalPP, tant et si bien qu'il est difficile de distinguer une part spécifique du budget et du temps alloué dans le cadre du projet à ces aspects.

La forge Gitlab de l'IPP lui est accessible à titre gracieux au nom du caractère universitaire de la fondation École d'économie de Paris dont il dépend.

Au-delà de la fin du projet Collex-Persée, l'Institut des politiques engage des ressources supplémentaires pour assurer la continuité de la base de données, c'est-à-dire pour assurer son entretien, sa mise à jour, et son accessibilité, ainsi que pour assurer l'animation de la communauté des contributeurs et utilisateurs de la base de données.

Liste des figures

L'Institut des politiques publiques (IPP) est développé dans le cadre d'un partenariat scientifique entre PSE-Ecole d'économie de Paris (PSE) et le Centre de Recherche en Économie et Statistique (CREST). L'IPP vise à promouvoir l'analyse et l'évaluation quantitatives des politiques publiques en s'appuyant sur les méthodes les plus récentes de la recherche en économie.

PSE a pour ambition de développer, au plus haut niveau international, la recherche en économie et la diffusion de ses résultats. Elle rassemble une communauté de près de 140 chercheurs et 200 doctorants, et offre des enseignements en Master, École d'été et Executive education à la pointe de la discipline économique. Fondée par le CNRS, l'EHESS, l'ENS, l'École des Ponts-ParisTech, l'INRA, et l'Université Paris 1 Panthéon Sorbonne, PSE associe à son projet des partenaires privés et institutionnels. Désormais solidement installée dans le paysage académique mondial, la fondation décloisonne ce qui doit l'être pour accomplir son ambition d'excellence : elle associe l'université et les grandes écoles, nourrit les échanges entre l'analyse économique et les autres sciences sociales, inscrit la recherche académique dans la société, et appuie les travaux de ses équipes sur de multiples partenariats. www.parisschoolofeconomics.eu

Le Groupe des écoles nationales d'économie et statistique (GENES) est un établissement public d'enseignement supérieur et de recherche. Au sein du GENES, le CREST est un centre de recherche interdisciplinaire spécialisé en méthodes quantitatives appliquées aux sciences sociales regroupant des chercheurs l'ENSAE Paris, de l'ENSAI, du département d'Économie de l'École polytechnique et du CNRS. http ://www.groupe-genes.fr/ – http ://crest.science