Les politiques d'accès aux ressources numériques de l'administration au prisme de la microsimulation

Franck Bessis, Paul Cotton

Les politiques d'accès aux ressources numériques de l'administration au prisme de la microsimulation[1]

Franck Bessis[2] et Paul Cotton[3]

Introduction

> « Pour vous raconter la petite histoire de la raison pour laquelle on s'est excité sur la microsimulation, c'est que Anthony Atkinson avait eu une expérience assez extraordinaire : il était relativement proche du Parti travailliste (…) Au moment du discours sur le budget au Parlement anglais, le premier ministre fait son discours et indique toutes les réformes qui vont avoir lieu. Atkinson avait convaincu Gordon Brown qu'il était possible de simuler pratiquement en même temps que le premier ministre allait faire son annonce les effets de ses réformes (…). Le premier ministre ce jour-là a fait son discours. Gordon Brown a demandé une interruption de séance d'une demi-heure, durant laquelle Atkinson a fait ses calculs sur son ordinateur qui était à l'époque rudimentaire, et il a sorti des résultats en disant « Voilà ce que va faire votre réforme, il y a tant de gagnants, tant de perdants, les gagnants sont ces gens-ci, les perdants sont ces gens-là ». Le Parlement anglais était absolument stupéfait de voir que l'opposition avait été capable, dans un temps aussi court, de produire ces résultats. Et c'est un petit peu là-dessus qu'on a lancé le projet de microsimulation, parce qu'on s'est dit qu'on allait convaincre les journalistes, les parlementaires d'utiliser systématiquement ce produit. » (Entretien avec François Bourguignon, juillet 2021).

Cet extrait d'entretien avec l'un des pionniers de la microsimulation en France résume bien l'enjeu d'une diffusion de cet instrument de conception et d'évaluation des politiques fiscales et sociales. Mettre des modèles de microsimulation à disposition d'un public élargi (chercheurs, journalistes, parlementaires, voire contribuables et/ou bénéficiaires de prestations sociales, électeurs) plutôt que d'en réserver l'usage aux seules administrations suppose également d'ouvrir l'accès aux données nécessaires à leur fonctionnement. La relation entre outils numériques et action publique peut alors être envisagée selon deux voies complémentaires dans le cas de cet objet. D'une part, en tant qu'instruments de réformes développés grâce aux progrès de l'informatique, les modèles de microsimulation se présentent comme des outils numériques pour l'action publique. D'autre part, la place prise par ces derniers au cours des trente dernières années invite aussi à s'intéresser à certaines formes d'action publique sur le numérique, à savoir la politique d'accès à des ressources numériques de l'administration, la manière dont des acteurs négocient des marges de manœuvre à partir de celles-ci en s'en emparant, en contournant son esprit, ou bien encore en cherchant à la faire évoluer. Cette action publique n'est donc pas simplement le produit d'une volonté politique, mais est le fruit d'une mobilisation d'acteurs de différents horizons : agents de l'Insee, chercheurs en économie, et militant de l'open-data au sein ou en dehors de l'administration. En prenant appui sur l'évolution du paysage de la microsimulation en France, ce chapitre se concentre sur deux enjeux de la politique d'accès aux ressources numériques de l'administration : le premier concerne l'accès aux modèles de l'administration (« open-access »), le second l'ouverture de ses données (« open-data »)[4]. Ces enjeux se traduisent par des affrontements plus ou moins directs avec l'administration, visent la promotion et l'expansion du renversement de tendance

[1] Chapitre 3 de Revest V. et Liotard I (dir.), *Transformation digitale et politiques publiques*, ISTE éditions, 2022.

[2] UMR 5206 Triangle et Université Lumière Lyon 2, franck.bessis@univ-lyon2.fr

[3] UMR 5206 Triangle et Science Po Lyon, paul.cotton@sciencespo-lyon.fr

[4] Sur le plan juridique, une base de données est entendue comme « un recueil d'œuvres, de données ou d'autres éléments indépendants, disposés de manière systématique ou méthodique, et individuellement accessibles par des moyens électroniques ou par tout autre moyen » (article L112-3 du 2 juillet 1998).

introduit par l'open-data, selon lequel « les données et informations publiques doivent être publiées en ligne avant même d'être réclamées par des tiers » (Beranger, 2017), ou autrement dit « la diffusion volontaire et proactive des données, essentiellement publiques, qui deviennent librement réutilisables » (Goeta, 2015).

Avant d'entrer plus en détail dans ce mouvement d'ouverture étudié ici au prisme de la microsimulation, il convient de présenter brièvement cet instrument. Les modèles de microsimulation peuvent être soit statiques, et reconstituer l'impact de la politique sur des individus à un instant donné ; soit dynamiques, et reconstituer cet impact tout au long d'un cycle de vie de l'individu (Blanchet, 2015). Ce chapitre se concentre sur les modèles statiques, appliqués au système de redistribution monétaire français. Il y sera plus particulièrement question des modèles Ines, Myriade, Openfisca, Saphir, et Taxipp[5]. Ces modèles sont constitués de plusieurs éléments. En premier lieu, ils reproduisent sous forme de code informatique la législation socio-fiscale sous une forme simplifiée[6]. Adossée à cette législation, une base de données micro-économique contenant des informations individuelles sur un échantillon représentatif de la population, permet de simuler les effets du système socio-fiscal en termes de redistribution et de masses budgétaires. Cette base peut être constituée à partir d'une enquête (par exemple, « Budget des familles »[7]) ou de l'appariement d'une enquête et de données administratives (par exemple, des déclarations fiscales et « l'enquête Emploi »[8]). A la différence de la législation socio-fiscale accessible à tous via le Journal Officiel[9], ces données ne sont pas en libre accès. Produites par les administrations elles-mêmes, une partie des informations qu'elles contiennent relèvent du secret professionnel (données relatives à la situation familiale ou personnelle, à la situation fiscale, etc.). Une fois ces deux ingrédients réunis (le code de la législation, et les données microéconomiques), les modèles de microsimulation permettent de modifier le système socio-fiscal en vigueur en paramétrant un ou plusieurs scénarios d'évolution. En comparant les résultats ainsi obtenus avec les résultats obtenus sans modification, l'évaluation des impacts redistributifs et budgétaires du ou des scénarios peut être réalisée. Ce sont également ces modèles qui servent à établir les bilans redistributifs dont la presse se fait largement l'écho chaque année autour de la présentation du projet de loi de finances[10].

Une première étude du mouvement d'ouverture des codes et données relatives à la microsimulation a été proposée par Shulz (2019). En se concentrant sur le processus d'élaboration d'Openfisca, l'auteur conclut que la réussite de la mobilisation en faveur de cette ouverture repose sur la « collaboration d'acteurs publics et non publics, à la périphérie de [l'] administration » (p.867). En resituant l'épisode d'Openfisca dans l'histoire du développement

[5] Pour un historique plus complet de la microsimulation statique en France, voir Legendre (2019) et Bessis et Cotton (2021).

[6] Ce code peut être écrit en différents langages informatiques. Par exemple, et de façon non exhaustive : SAS, pour un langage qui nécessite une licence payante ; C++, R, ou Python pour des langages « libres », et donc gratuits.

[7] https://www.insee.fr/fr/metadonnees/source/serie/s1194

[8] https://www.insee.fr/fr/metadonnees/source/serie/s1223

[9] Le coût d'entrée pour sa compréhension reste cependant élevé.

[10] Dernier exemple en date : le bilan redistributif du quinquennat d'Emmanuel Macron produit à partir du modèle de microsimulation de la Direction générale du Trésor pour le Rapport économique, social et financier 2022 annexé au projet de loi de finances. « Macron, président des riches. L'heure des comptes », *Libération,* 11 octobre 2021, pp.1-5 ; « Macron cherche à effacer son image de "président des riches" », *Le Monde*, 6 octobre 2021, p.16 ; « L'exécutif satisfait d'avoir augmenté le pouvoir d'achat des Français depuis 2017 », Le Figaro, 5 octobre 2021, p.22 . Suivi des contre-chiffrages produits quelques semaines plus tard par l'Institut des Politiques Publiques : « Le quinquennat Macron a bien été celui des ultras-riches », Médiapart, 17 novembre 2021 ; « Pouvoir d'achat : l'étude qui relance le débat sur les gagnants du quinquennat », Les Echos, 17 novembre 2021, p.2 ; « Pouvoir d'achat : le mandat Macron à la loupe », Le Monde, 17 novembre 2021, p.14.

des modèles de microsimulation statique, nous proposons de compléter cette analyse sur deux aspects. Le premier concerne l'intervention d'autres acteurs situés au cœur de l'administration, et plus précisément au cœur du système de statistique publique. Le second porte sur les divergences de vues entre les acteurs qui ont pris part à ce mouvement : plutôt qu'une « collaboration d'acteurs », nous entendons faire apparaître et mettre en débat une pluralité de conceptions de la qualité de l'ouverture, c'est-à-dire à la fois des conceptions de ce qu'il importe d'ouvrir et de comment bien ouvrir.

Le concept privilégié dans cette optique est celui de « statactivisme », qui vise l'ensemble des actions menées de manière à « mettre les statistiques au service de l'émancipation » (Bruno, Didier et Prévieux (2014)). Les demandes d'ouverture des codes et données de l'administration se situent en amont des manières de « faire de la statistique une arme critique » étudiées jusqu'à présent dans cette perspective, qu'il s'agisse de consolider de nouvelles catégories « sur lesquelles s'appuyer pour revendiquer des droits et défendre des intérêts » (par exemple la catégorie d'*intellos précaires*) ou « opposer des indicateurs alternatifs à l'institution » (comme les *nouveaux indicateurs de richesse*). Avant de « lutter avec des nombres » les acteurs dont nous restituons les démarches luttent pour l'accès aux moyens de production de l'institution. L'une de ces formes originales de « stratégie statactiviste » (celle qui vise l'accès aux données) a des conséquences directes sur l'affirmation d'une catégorie collective revendiquée sur les pancartes de manifestants au tournant des années 2010 avec le slogan « Nous sommes les 99% ». L'autre stratégie centrée sur l'accès aux modèles économiques de l'administration vise moins une définition alternative de la réalité qu'une transformation des conditions du débat démocratique, en dotant les acteurs de ce débat de nouvelles capacités d'argumentation et de contre-argumentation[11].

Ce chapitre vise donc précisément à décrire la façon dont ces deux stratégies s'incarnent, se développent et cohabitent dans l'histoire de la microsimulation[12]. L'accès aux données microéconomiques et l'accès aux codes des modèles, aussi bien entre administrations qu'en dehors, constituent en effet deux phénomènes bien différents, portés par des « entrepreneurs de cause » (Cobb, Elder, 1972) aux profils et motivations souvent distincts. Au-delà de la question de savoir pourquoi certains acteurs portent leur attention sur l'accès aux données plutôt qu'aux codes (ou inversement), ce chapitre cherchera à apprécier l'efficacité des différentes stratégies déployées pour y accéder. L'analyse repose sur le croisement de deux sources de données empiriques. Une étude de la littérature scientifique sur le sujet a été réalisée pour mieux identifier l'écosystème de la microsimulation, ses acteurs, et ses enjeux Cet usage est justifié par le fait que la littérature académique sur la microsimulation a principalement été écrite par des acteurs investis dans le développement de cette pratique ; et dont la plupart ont été rencontrés lors d'entretiens. Une quarantaine d'entretiens a également été réalisée entre 2019 et 2021 avec des acteurs investis dans la microsimulation, principalement des chercheurs et

[11] C'est dans ce sens que nous comprenons l'anecdote rapportée au début de cette introduction. On peut également mentionner à titre d'exemple les effets attendus de la production au début des années 2000 d'une mesure alternative de l'évolution des inégalités avec le Baromètre des inégalités et de la pauvreté (BIP 40) : « l'intérêt d'un baromètre comme le BIP 40 n'est pas seulement d'apporter des éléments de constat sur les inégalités et la pauvreté, mais aussi de permettre à l'ensemble des acteurs concernés d'exercer leur réflexion critique, de débattre et de mieux agir pour combattre ces inégalités » (Concialdi (2014), p.211).

[12] Ces deux stratégies ne sont toutefois pas exhaustives, puisqu'une partie des acteurs œuvrent pour une réutilisation des codes et des données en dehors du seul cadre du débat démocratique. Voir notamment datactivist.coop, société coopérative et participative (SCOP) fondée entre autres par Samuel Goeta suite à ses travaux de thèses portant sur l'open data.

hauts-fonctionnaires, mais aussi un représentant de l'association qui a forcé l'ouverture de plusieurs modèles.

Dans une première partie, nous présenterons la situation qui a prévalu jusqu'au début des années 2000. A cette période, le développement de la microsimulation en dehors de l'administration est limité par un cadre législatif peu favorable à l'utilisation par les chercheurs des données fiscales. Ce cadre va évoluer progressivement, permettant un accès possible mais non obligatoire aux données les plus adaptées à la simulation du système socio-fiscal. Dans une deuxième partie, nous tâcherons de montrer pourquoi et comment émerge à partir des années 2010 un mouvement d'ouverture de ces données. Une troisième partie portera sur l'ouverture des codes de l'administration, qui suit de près celui des données. Enfin, la quatrième partie sera l'occasion de confronter les conceptions de la qualité de l'ouverture des codes et données défendues par les différents acteurs de notre histoire.

1. D'une fermeture contournée à une ouverture progressive et non-systématique des données (1951-2001)

L'histoire de l'ouverture des données des administrations françaises et de leurs conditions d'accès[13] a largement été documentée, notamment s'agissant de bases de données statistiques et d'enquête (Rhein, 2002 ; Chenu, Silberman 2011 ; Caporali et *al.*, 2015). Le développement de la microsimulation s'inscrit pleinement dans cette histoire, avec quelques spécificités sur lesquelles nous nous centrons par la suite afin de mieux saisir le terreau dans lequel le double mouvement d'ouverture (des codes et des données) s'est ensuite incarné.

1.1. En dehors de l'administration, des chercheurs qui parviennent à accéder à des données de façon « officieuse »

Les premières expériences de microsimulation sont conduites en France à partir de la fin des années 1960 au sein de la Direction de la Prévision. Les universitaires emboîtent le pas quelques années plus tard, avec le modèle Sysiff développé dans la décennie 1980 au sein du laboratoire Delta (EHESS). Si l'on comprend naturellement que la question de l'accès aux données micro-économiques nécessaires au fonctionnement de leur modèle ne présente, à l'époque, aucune difficulté pour la Direction de la Prévision, elle se pose en revanche pour les équipes du Delta. Et pour cause, à cette période, le principal cadre juridique de référence pour les chercheurs est celui instauré par la loi de 1951 sur le secret statistique[14]. De ce fait, les bases de données qui contiennent des informations relatives à la situation personnelle, familiale ou privée des individus ne peuvent être communiquées en dehors de l'administration, et ce même à des fins de recherche. Ces données sont en effet considérées comme « confidentielles », et ne doivent pas circuler. Autrement dit : les chercheurs ne peuvent demander l'accès à aucune des données d'enquêtes de l'administration. L'esprit de cette loi est en effet d'abord d'assurer « l'étanchéité entre les services de l'Etat » pour protéger et maintenir la confiance des personnes sollicitées pour ces enquêtes, en garantissant que leurs informations (c'est-à-dire *leurs données*) ne seront utilisées qu'à des fins statistiques - et non par exemple, pour vérifier s'ils ont bien rempli leur déclaration d'impôt (Silberman, 2011). Une évolution législative intervient dans le courant de

[13] Les dispositions législatives permettent et encadrent l'accès aux données sont regroupées au sein du Code des Relations entre le Public et l'Administration, entrée en vigueur au 1er janvier 2016 dans la continuité du « choc de simplification » voulu par le Président François Hollande.

[14] Loi 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques.

l'année 1978[15]. Cette dernière offre la possibilité aux administrés (et donc aux chercheurs) de demander la consultation de tout document administratif (y compris des données statistiques) à condition que ces derniers ne soient pas nominatifs ou indirectement nominatifs[16]. L'accès aux données d'enquête de l'administration sur demande est donc possible en théorie. Cependant, il ne s'agit pas d'un accès par défaut. En effet, l'administration peut refuser l'accès pour un certain nombre de motifs, en invoquant par exemple « un secret protégé par la loi » (article 6 de la loi de 1978). Dans ce cas, le demandeur doit saisir la Commission d'Accès aux Documents Administratifs (CADA), nouvellement créée, qui doit alors formuler un avis sur le caractère fondé ou non du refus. Il convient ensuite aux demandeurs de négocier à nouveau l'accès au document administratif, puis, en cas de nouveau refus de porter sa demande auprès du Tribunal administratif.

Alors que l'on pourrait s'attendre à ce que les équipes qui s'attèlent au développement du modèle Sysiff tentent d'accéder aux données d'enquête de l'administration en s'appuyant sur la loi de 1978, ces chercheurs ont utilisé des données obtenues de deux autres façons. La première semble être[17] la réutilisation d'une source mise à disposition pour un précédent contrat de recherche. Pour sa première version, Sysiff utilisait ainsi une base de l'Insee - l'Enquête Revenu Fiscaux (ERF), à laquelle ses concepteurs avaient eu accès grâce à une recherche menée pour la Caisse nationale des allocations familiales (Cnaf). Cette pratique n'était pas considérée comme problématique car non visible, pour peu que les chercheurs n'en fassent pas la publicité[18]. La seconde façon d'accéder aux données tient dans la transmission directe, par voie non officielle. A l'époque, la diffusion des données est laissée de façon tacite à la discrétion des personnes qui y avaient accès au sein de l'administration : soit directement des responsables d'enquête, soit des personnes qui exploitent ces bases à des fins statistiques, soit des personnes situées à un niveau hiérarchique plus élevé. Les créateurs du modèle Sysiff ont de cette façon eu accès aux données d'une autre enquête de l'Insee (Budget des familles). Cette transmission impliquait un travail d'anonymisation pour s'assurer que les données transmises ne soient pas directement ou indirectement nominatives. Et à nouveau, cette pratique ne créait pas de difficulté particulière tant qu'elle restait discrète.

> « A l'époque relativement peu de chercheurs travaillaient avec des données microéconomiques parce qu'elles n'étaient pas disponibles. J'ai fait des travaux en Syldavie[19] sur des bases de données microéconomiques et je dois dire que les bases de données en question je les ai purement et simplement volées, passées en contrebande, en sortant de Syldavie (…). Je revenais avec ces énormes bandes d'ordinateurs qui faisaient rêver à l'époque, on disait « voilà ça c'est l'an 2000 ». Je sortais de Syldavie avec trois ou quatre bandes de ce type-là sous le bras. Un jour à Paris, j'ai été voir le directeur de l'Insee de l'époque (…) et je lui ai demandé « Comment se fait-il qu'en France les chercheurs n'ont pas la possibilité d'utiliser ces données microéconomiques ? » Et très discrètement il m'a fait donner l'enquête Budget des Familles, là aussi pratiquement en contrebande, en me disant « On vous la donne, on va l'anonymiser un tout peu plus, mais n'en parlez pas trop » (…). Il ne voulait pas non plus prendre ouvertement et publiquement l'engagement de dire « à partir de maintenant, on va mettre ces données-là en accès public ». » (Entretien avec François Bourguignon, juillet 2021).

[15] Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.

[16] Cette expression vise la possibilité d'identifier des personnes à partir d'un croisement de variables.

[17] Ce premier élément n'a pas été directement confirmé par les enquêtés concernés qui ne se souvenaient plus des conditions exactes dans lesquelles ils avaient eu accès à l'ERF tout en admettant que cette interprétation était probable, le rapport pour la Cnaf faisant bien état d'une utilisation de l'ERF peu avant la création de Sysiff.

[18] Une autre ré-utilisation de l'ERF indépendante de Sysiff a fait l'objet d'une publication dans une revue académique.

[19] Le nom du pays a été modifié.

On retrouve ici illustré par les premières phases de l'histoire de la microsimulation, le constat général dressé par Caporali et *ali* (2015, p.575) : « aucun droit d'accès des chercheurs universitaires aux données produites par le service statistique public n'était prévu par les accords passés avec le CNRS et l'Insee, ce qui conduisait certains d'entre eux à s'appuyer sur des contacts personnels au sein de l'Insee ou d'autres administrations » (*Ibid*, p.575). La situation va progressivement évoluer vers des relations plus formelles, via la mise en place de conventions entre le CNRS et l'Insee (Silberman, 2011). La première sera signée en 1986, et sera renouvelée jusqu'à la mise en place du Réseau Quetelet en 2001. Ce dernier aura la charge de rendre accessible aux chercheurs des données de sciences sociales (dont des données d'enquêtes des ERF 1970 à 1990) pour peu que leur usage soit limité à des fins de recherche et sous couvert d'anonymisation des fichiers (et donc sans certaines données permettant de reconstituer indirectement l'identité des enquêtés).

Ces façons d'accéder aux données posent toutefois un certain nombre de limites pour le développement des modèles de microsimulation en dehors de l'administration.

En particulier, bien qu'ils se soient engagés parmi les premiers dans la microsimulation, les chercheurs n'ont pu rivaliser avec les modèles développés en son sein (Ines, Myriade, puis Saphir). Un écart qui tient, entre autres choses, à la nature des données disponibles.

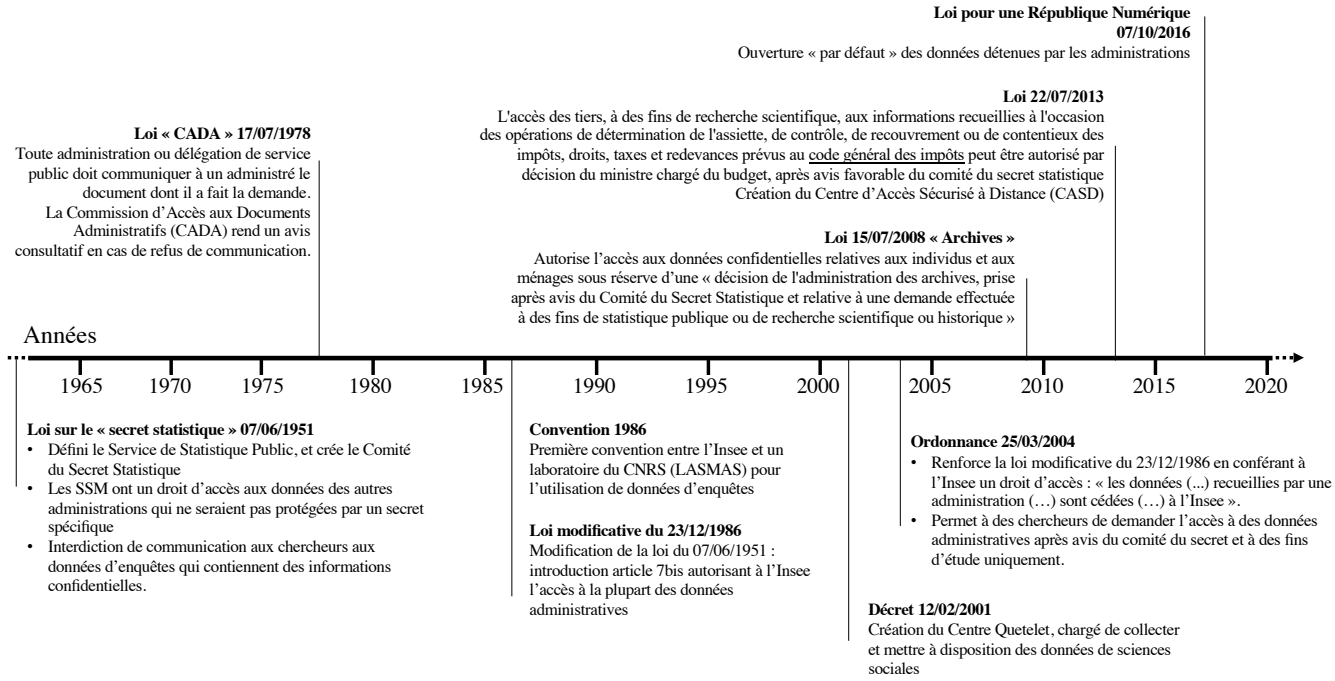

Figure 1. Les grandes étapes de l'ouverture des données des administrations dans l'histoire de la microsimulation.

1.2. Un accès incomplet : l'avantage décisif des économistes de l'administration

Les données à la disposition des chercheurs entre les années 1980 et 2000 peinent à rivaliser avec celles qu'utilisent les administrations engagées dans la microsimulation, et ce sur plusieurs aspects. Pour le comprendre, une précision s'impose sur la nature des données utilisées dans ces modèles. Pour rappel, un modèle de microsimulation nécessite une base de données individuelles sur laquelle appliquer les différents scénarios à comparer. Plusieurs types de sources ont été utilisées au cours des différentes expériences de microsimulation en France : (i) des données d'enquête telles que l'enquête Budget des Familles ; (ii) des données

administratives : des données issues des déclarations fiscales, échantillonnées comme le fichier échantillonné de l'impôt sur le revenu (FELIN) ou exhaustive comme le fichier POTE ; (iii) des appariements entre données d'enquête et données administratives : l'enquête revenus fiscaux (ERF) résulte d'un appariement des données de l'enquête emploi avec le fichier des déclarations fiscales. Ces sources, produites par différentes administrations, contiennent des données qui leur sont propres, et qui permettent donc de simuler avec plus ou moins de précisions (sur plus ou moins d'aspects, plus ou moins d'individus ou de ménages). De façon générale, trois critères permettent d'apprécier leur « qualité » relative. D'abord, les taux de sondage appliqués selon les catégories de population que l'on souhaite examiner plus finement (l'exhaustivité dans le haut de la distribution étant même devenu un enjeu de première importance pour les microsimulateurs au cours des deux dernières décennies). Ensuite, leur périodicité : est-ce que la base est actualisée à intervalle régulier, tous les ans (comme aujourd'hui pour l'ERF), tous les deux ans, tous les quatre ans (ancienne périodicité de l'ERF avant 1996) ? Les modèles s'appuient-ils sur le dernier millésime, ou sur une version datée ? Enfin, leur nature : s'agit-il de données d'enquête ou de données administratives. De ce fait, toutes les données ne se valent pas.

Cela étant posé, on comprend les limites inhérentes aux façons d'accéder aux bases qui ont prévalu jusqu'au début des années 2000. Comme le signale Legendre (2019), « pas de modèle de microsimulation sans un échantillon représentatif suffisamment fiable », c'est-à-dire se rapprochant des critères de qualité relative évoqués. Au début des années 2000, Sysiff utilise la source Budget des Familles basée sur des données d'enquête. Un autre modèle développé par des chercheurs de l'Université de Cergy en collaboration avec l'Observatoire français des conjonctures économiques (Ofce) au début des années 2000 bénéficie en revanche de millésimes plus récents (ERF 1998, Budget des familles 2000 et 2001, obtenues grâce au contrat passé avec le Parlement pour le développement du modèle). S'il peut un temps jouer sur le même plan que les modèles de l'administration (et en particulier celui de l'Insee, qui utilise également ERF pour son modèle de microsimulation Ines), la non-actualisation des bases[20] va contribuer à rendre le modèle de moins en moins en capacité de produire des résultats adaptés à l'année en cours. De plus, il ne suffit pas de disposer d'une meilleure base pour produire un meilleur modèle, puisque l'intégration d'une nouvelle base de données qui ne serait pas une simple actualisation de la précédente nécessite un travail de (re)développement du modèle. Quand bien même les chercheurs auraient eu accès à ces données en dehors de l'administration, un travail de développement aurait ainsi dû être assumé pour intégrer ces évolutions dans les modèles en question.

Cet enjeu d'accès aux données se retrouve aussi, dans une moindre mesure, au sein des différents membres du Service de statistiques public (SSP, composé des services de statistiques ministériels[21] et de l'Insee) engagés dans la microsimulation. Au tournant des années 1990, malgré l'expérience du modèle Mir au sein de la Direction de la prévision une vingtaine d'années plus tôt, la microsimulation statique en est encore à ses débuts au sein du Service de statistique public (SSP). Les données sont alors un moyen pour l'administration de se

[20] Depuis 1996, l'Enquête Revenus Fiscaux est mise à jour annuellement, contre 5 ans auparavant. Des nouvelles versions sont ainsi produites chaque année.

[21] L'annexe du décret décret n° 2009-250 du 3 mars 2009 relatif à l'Autorité de la statistique publique, mis à jour par arrêté le 11 septembre 2020, définit 16 services de statistiques ministériels, répartis sur l'ensemble des champs ministériels. On y trouve par exemple la Direction de l'animation de la recherche, des études et des statistiques (Dares) côté Travail ; la Direction de la recherche, des études, de l'évaluation et des statistiques (Drees) côté Santé et Solidarités ; ou encore le Pôle statistique publique, division du département des études et statistiques fiscales (DESF) de la Direction générale des finances publiques (DGFiP) côté Économie et Finances. L'ensemble des SSM est coordonné par l'Insee.

positionner et de s'affirmer comme acteur spécifique et légitime dans le paysage de la microsimulation, et en particulier par rapport aux chercheurs externes. La plupart de ces administrations sont en effet productrices des données microéconomiques utiles à la microsimulation, et bénéficient donc d'un accès direct aux données les plus actualisées. L'Insee disposait pour son modèle baptisé Ines (développé à partir de 1996) des dernières versions de l'ERF puisque produites en interne. Bien que ne faisant pas partie du SSP, le service statistique de la Cnaf dirigé par un administrateur de l'Insee a, pour son modèle Myriade, également eu accès sans difficulté à l'ERF. La Direction de la Prévision jouit quant à elle pour son modèle Saphir (développé à partir du début des années 2000) d'un accès à l'échantillon dit « lourd » de la Direction générale des impôts (équivalent de la base « FELIN », qui contient 500 000 déclarations fiscales). En un sens, la Direction de la Prévision dispose alors d'un avantage sur les modèles de l'Insee et de la Cnaf en raison de la nature des données utilisées. Elle a la capacité d'observer plus finement les évolutions de revenus d'une année sur l'autre et sur une période plus récente. Dans une certaine mesure, cette asymétrie peut être lue comme une façon pour ces administrations de conserver des ressources propres, mobilisables pour s'affirmer dans l'expertise en matière d'évaluation des dispositifs sociaux et fiscaux, à la fois vis-à-vis des chercheurs externes, mais aussi vis-à-vis des autres administrations (Pénissat, 2009). Néanmoins, il convient de rappeler que l'Insee aurait pu utiliser dans son modèle l'exhaustif des données fiscales, auxquelles elle peut accéder en théorie depuis 1951, dans les faits depuis 1986[22], et a accédé pour constituer l'appariement de l'ERF (Legendre, Lorgnet, Thibault, 2001). Le véritable écart sur le plan des données se situe ainsi bien entre modèles internes et externes à l'administration. En effet, les données les plus récentes dont disposent les chercheurs sont, pour la plupart du temps, uniquement des données d'enquête ; là où les membres du SSP engagés dans la microsimulation disposent à l'époque a minima de données d'enquête appariées à des données administratives directement issues de déclarations fiscales, et non de déclarations d'enquêtes[23]. Les membres du SSP utilisent également des données actualisées, là où les chercheurs doivent à l'époque se contenter de données moins récentes.

Jusqu'au début des années 2000, les chercheurs usent ainsi de relations informelles plutôt que de l'affrontement pour avoir accès aux données de l'administration. La faiblesse du cadre juridique de l'époque, peu connu à la fois des chercheurs et des agents, permet « une plus grande souplesse dans la mise en circulation [des données d'enquêtes] en particulier auprès des chercheurs » (Pénissat, 2009). Ces derniers doivent cependant se contenter la plupart du temps des seules données d'enquêtes, même si l'obtention de contrat de recherche ou des conventions avec des membres du SSP permettent ponctuellement un accès à des données d'enquêtes

[22] Dans sa version initiale, la loi de 1951 octroyait à l'Insee et aux SSM un droit d'accès à des fins de production statistique aux données administratives des autres administrations qui ne seraient pas couvertes par un secret légal. Ce droit visait à alléger le travail d'enquête en utilisant des informations déjà récoltées par d'autres administrations. Les données récoltées par l'administration fiscale étaient toutefois des données protégées (article 2006-2 du CGI de l'époque). Il faut attendre l'introduction d'un article "7 bis" à la loi de 1951 via la loi n° 86-1305 du 23 décembre 1986 et l'article L. 135D du livre des procédures fiscales pour résoudre la situation : « Les informations relatives aux personnes physiques (...) recueillies (...) par une administration (...) peuvent être cédées, à des fins exclusives d'établissement de statistiques, à l'Institut national de la statistique et des études économiques (…) les dispositions de l'alinéa précédent s'appliquent nonobstant toutes dispositions contraires relatives au secret professionnel. ». S'il s'agit bien là d'une possibilité et non d'un droit ("peuvent être cédées", et non "doivent être cédées), l'Insee peut désormais accéder aux données sans que les agents du ministère de l'économie et des finances se mettent en contradiction avec le secret fiscal.

[23] L'enquête Budget des familles prévoit par exemple que les personnes interrogées tiennent un carnet de comptes pendant plusieurs jours pour noter toutes leurs dépenses (volet consommation) et déclarer leur revenu (volet revenu). Des erreurs ou oublis peuvent ainsi se glisser dans la déclaration faite par les interrogés, impactant directement la qualité des données recueillies.

appariées à des données administratives (ERF)[24]. Largement tolérées par les administrations concernées, ces pratiques vont finalement se retourner contre les chercheurs lorsque certains d'entre-eux vont franchir une « ligne rouge » à partir des années 2010 : la réutilisation des données à des fins politiques pour participer au débat public (voir *infra*). Cet épisode ouvrira cependant la voie vers davantage de formalisation dans l'accès aux données pour les chercheurs, dont la forme la plus avancée à ce jour est la Loi pour une République Numérique adoptée en 2016.

	BDF	**ERF**	**POTE**	**FELIN**
	Données d'enquête			
		Données administratives (déclarations fiscales)		
Sysiff - Laboratoire Delta	**X**			
Ines - Insee-Drees-Cnaf		**X**		
Myriade - Cnaf		**X**		
Saphir - DG-Trésor		**X**		
OpenFisca - France Stratégie et Etalab		**X**		
Taxipp - Institut des Politiques Publiques		**X**	**X**	**X**

Tableau 1. Récapitulatif des sources de données utilisées par chaque modèle de microsimulation.

2. Le mouvement d'ouverture des données au tournant des années 2010 : du repli au changement institutionnel

L'ouverture des données de l'administration en dehors des agents des SSM suit une trajectoire particulière. Le premier modèle « grand public » du début des années 2010 utilise des données originales au regard des précédentes expériences de microsimulation, mais ces données sont détournées à la fois de leur finalité première et des usages autorisés. Alors que cet épisode a ravivé la prudence des administrations vis-à-vis des chercheurs concernant la transmission de données sensibles, plusieurs initiatives vont par la suite amplifier l'ouverture à partir du milieu des années 2010.

2.1. Révolution fiscale et contre révolution statistique : un mouvement de fermeture des données

Dans la deuxième partie des années 1990, pour étudier les hauts revenus en France, Thomas Piketty accède grâce à un contrat avec la Direction de la Prévision à des sources encore difficilement accessibles au chercheur à l'époque : les échantillons de déclarations de la DGI de 1988 à 1995 et les ERF de 1970 à 1990 (Piketty, 1998). Sa participation au rapport du Conseil d'analyse économique sur les inégalités économiques paru en 2001 lui donne une nou- velle occasion de travailler sur ces données, tout en formulant le souhait d'ac- céder aux données fiscales exhaustives de la DGI pour mieux rendre compte des évolutions de la situation des plus fortunés. Quelques années plus tard, alors qu'il prépare une thèse sous sa direction, Camille Landais peut s'appuyer sur les données fiscales exhaustives produites par l'administration fiscale pour actualiser les séries de Piketty (couvrant la période 1970-1996) sur la période allant

[24] Dans d'autres domaines, certains chercheurs parviennent toutefois à accéder à des données administratives en usant de relations informelles. Voir par exemple les conditions d'accès des chercheurs aux bases de la Dares (Pénissat, 2009, pp.403-404).

de 1998 à 2005 (Landais, 2007). L'accès à ces données permet d'explorer la distribution des revenus à un niveau de détail plus fin, ce qui concentre l'attention sur de nouvelles catégories d'acteurs (les 1% les plus riches) et montre une croissance rapide des inégalités sur la période étudiée par C. Landais. Ce résultat conforte les acteurs qui, s'étonnant du diagnostic jusque-là établi par l'Insee d'inégalités à peu près stables d'une année sur l'autre, plaidaient à la même époque pour une nouvelle approche des inégalités[25].

A cette première forme de statactivisme (affirmation d'une catégorie collective), Piketty et Landais en ajoute bientôt une seconde (transformation des conditions du débat démocratique), en donnant l'accès au grand public à un nouveau modèle de microsimulation créé pour l'occasion pour permettre à chacun de simuler sa propre réforme fiscale[26]. Le site accompagne la sortie de l'ouvrage *Pour une révolution fiscale* (auquel participe aussi Emmanuel Saez) dont la parution à un an de l'élection présidentielle de 2012 (où l'étiquette de « président des riches » colle déjà au président sortant) ne passe pas inaperçue. Pour alimenter leur modèle, les auteurs s'appuient à nouveau sur des données fiscales et notamment l'ERF 2006[27]. Les conditions dans lesquelles l'ensemble de ces données ont été obtenues n'ont pas pu être complètement éclaircies. L'hypothèse la plus probable semble être celle d'une réutilisation de données mises à disposition pour d'autres projets[28]. Interdite par les conventions de recherche, qui engagent les chercheurs à supprimer les fichiers au terme du projet pour lesquelles ils ont été accordés, la pratique consistant à réutiliser des données sans demander de nouvelles autorisations n'étaient toutefois pas exceptionnelles à l'époque. Surtout, cette pratique pouvait d'autant plus être tolérée que la diffusion des résultats produits restait limitée à la communauté académique. Quelles que soient les conditions exactes dans lesquelles les trois chercheurs ont eu accès à l'ERF pour ce projet, cette utilisation politique et bien plus médiatisée des données a provoqué une réaction de repli de l'administration.

Cette situation va compliquer l'accès aux bases pour les chercheurs qui ont souhaité, à partir de 2011, poursuivre sous le nom de Taxipp le développement d'un nouveau modèle de microsimulation au sein de la Paris School of Economics[29]. Ceux-ci se sont vu refuser de plus en plus de demandes de conventions d'accueil (prévues par la loi de 1978) pour utiliser les données produites par des administrations directement dans leurs locaux. Cela s'avérait d'autant plus problématique que, si l'équipe de *Révolution fiscale* disposait de plusieurs sources de données administratives (en plus d'ERF, le modèle reposait sur plusieurs données d'enquêtes : Budget des ménages, Patrimoine, Logement, Emploi…[30]), les équipes suivantes n'avaient accès qu'à l'agrégation globale de ces bases, et non les bases brutes, rendant complexe tout nouveau développement du modèle.

---

[25] Sujobert (2014) raconte la manière dont cette critique, portée au niveau du Conseil national de l'information statistique a conduit à la création d'un groupe de travail actif entre mi-2004 et janvier 2007 et participé à faire évoluer l'Insee sur ces questions, en rendant visibles les derniers centiles sous le dernier décile.

[26] Voir www.revolution-fiscale.fr/simulez-votre-propre-reforme-fiscale

[27] L'enquête ERF est devenue l'Enquête Revenus Fiscaux et Sociaux (ERFS) à partir de 2005. Par souci de lisibilité, nous conserverons dans la suite de ce chapitre l'appellation originelle "ERF"

[28] Outre les projets de recherche déjà évoqués plus haut, on peut également mentionner un mémoire de fin d'étude soutenu à Paris School of Economics en 2009 pour lequel l'auteur avait eu accès à l'ERF 2006.

[29] Le modèle Taxipp développé par l'Institut des politiques publiques (rattaché à la PSE) repart des programmes développés pour *Révolution Fiscale* (modèle rebaptisé après-coup Taxipp v.0.0).

[30] Pour une présentation détaillée des fichiers source, voir l'annexe technique disponible en ligne sur le site de *Révolution fiscale*.

2.2. D'une relation de méfiance à une relation de confiance : l'IPP et la LPR/Loi Lemaire (2011-2016)

C'est le blocage auquel est confrontée l'équipe en charge du développement du modèle Taxipp qui va paradoxalement être moteur dans l'accélération du mouvement d'ouvertures des données micro-économiques. Alors dirigée par Antoine Bozio (directeur de l'Institut des politiques publiques dont le projet initial est le développement du modèle Taxipp), l'équipe se voit refuser l'accès aux données de l'enquête ERFS par la DGFiP suite à se demande effectuée fin 2011[31]. Le secret imposé par le Livre des procédures fiscales est invoqué, quand bien même des accès périodiques avaient pu être attribués auparavant. Avec le soutien du directeur du Comité du secret, et du cabinet du Premier ministre de l'époque et en particulier de son conseiller économique Fabien Dell qu'A. Bozio a côtoyé dans le cadre de précédents travaux de recherche (Schulz, 2019), un processus de réforme s'enclenche pour clarifier le droit concernant l'accès à ces données. Les services du Premier ministre aboutissent finalement à un projet de loi accordant l'accès à des fins de recherche et d'évaluation. Ce projet a dû être retravaillé à plusieurs reprises : mésinterprétation des objectifs de la loi par les services en charge de sa rédaction, reprises par les juristes du SGG pour être en conformité avec le droit européen (ne pas créer un monopole avec le Centre d'accès sécurisé aux données - CASD), projet retoqué comme « cavalier budgétaire » dans le projet de loi de finance de 2012… Il faut attendre 2013 pour que le projet soit finalement intégré dans la loi sur l'enseignement supérieur et la recherche, puis ses décrets d'application en 2014, pour que les équipes de l'IPP aient accès, à partir de mars 2015, aux données précédemment couvertes par le secret fiscal : ERF, mais aussi l'échantillon « lourd » des déclarations d'impôt sur le revenu. Ces données ont été déterminantes dans la conception de la version 1.0 de Taxipp.

L'accès aux données couvertes par le secret fiscal permis, s'est ensuite posée la question de l'accès aux données de la Sécurité sociale. En 2016, Antoine Bozio explique avoir essuyé un refus lors de sa demande d'accès aux données de l'échantillon national des allocataires de la Cnaf, alors même que ce dernier réalisait avec ses équipes un travail d'évaluation pour la Cour des comptes, et qu'il y avait eu accès deux ans plus tôt. La rupture du secret professionnel est invoquée. Malgré le soutien du directeur du pôle politiques sociales de Matignon, qui a réussi à obtenir un accord de la direction générale de la Cnaf pour l'ouverture des données, les services n'avaient toujours pas transmis les données plusieurs mois après. Finalement, un nouvel article de loi a été élaboré pour « garantir des conditions de sécurité juridique pour les administrations qui autoriseraient l'accès à leurs données ». L'article est inséré dans la loi pour une République Numérique dite « loi Lemaire », adoptée fin 2016. Ce dernier prévoit ainsi que les administrations ne sont pas en tort si elles transmettent ces données à des chercheurs. Il n'offre ainsi pas un droit d'accès au chercheur : les administrations « peuvent », et non ne « doivent »[32]. Cependant, il s'avère qu'une administration refuse rarement l'accès aux données aux chercheurs qui souhaiteraient réaliser des études sur l'impact de ses politiques publiques, puisque le refus est difficilement à assumer publiquement : « Si la loi prévoit que "les administrations peuvent communiquer les données", cela implique bien qu'elles gardent le contrôle de la communication des données. Mais, dans la pratique, un équilibre est trouvé, et elles donnent généralement leur accord »[33].

[31] Dans cette partie, nous nous appuyons principalement sur un entretien réalisé avec Antoine Bozio en juillet 2021 et une annexe de son mémoire d'HDR (Bozio, 2018) dans laquelle il présente les démarches entreprises pour accéder aux données utilisées pour Taxipp (pp. 63-68).

[32] Les conditions de partage des données entre administrations se trouvent également facilitées.

[33] Entretien avec Antoine Bozio, juillet 2021.

Au-delà de ces évolutions, le rapport du Cnis co-écrit en 2017 par A. Bozio et P.-Y. Geoffard à la demande de la Secrétaire d'État au numérique dans le prolongement de la loi LPR, fait état de plusieurs difficultés persistantes dans l'accès aux données administratives à des fins de recherche. En particulier, la LPR prévoit l'accès à la plupart des données, mais pas leur traitement, et pas toujours leur appariement. Seules les données déjà utilisées à des fins de traitement statistiques au sein de l'administration peuvent faire l'objet d'un traitement par les chercheurs[34] . Les données utilisées à des fins de gestion au sein des administrations ne peuvent ainsi pas faire l'objet d'un traitement. Le rapport fait également état d'une large mésinterprétation de la loi par les administrations. Par exemple, les injonctions de la Cnil pour la non-conservation des données après leur utilisation étaient interprétées comme des injonctions à la destruction des données.

Si tout n'est pas encore parfait, les initiatives entreprises au début des années 2010 ont permis aux chercheurs et aux administrations de nouer entre eux des relations de confiance mutuelle, balayant l'image de chercheurs « non fiables », redevenue prégnante à la fin des années 2000. La construction d'un cadre juridique progressif jusqu'en 2016, et en particulier la mise en place d'un centre d'accès sécurisé à distance (CASD) en 2013 ont permis de rassurer les administrations craintives de communiquer des données couvertes par un secret. Cette transformation institutionnelle réussie contraste avec la manière, plus frontale, dont les codes des modèles ont été ouverts, pour laquelle « l'administration avait tous les moyens de ne pas faire ».

3. Le mouvement d'ouverture des codes : libre consenti *versus* libre forcé

Disposer de données micro-économiques est indispensable pour faire tourner les simulations d'un modèle. Le code du modèle, en particulier celui de la législation socio-fiscale peut être reconstitué à partir de la loi, mais pas les données micro-économiques. Ne pas avoir accès aux codes des modèles ne devrait ainsi pas être bloquant pour une administration ou une équipe de chercheurs qui disposerait des bases de données, et qui souhaiterait se lancer dans la microsimulation. Pourtant, disposer de ce code peut faire gagner un temps précieux, puisqu'il permet d'économiser une partie du temps de développement du modèle (reproduire la législation sous forme de code). Un temps qui nécessite de nombreux mois, notamment pour identifier puis transcrire la législation socio-fiscale en vigueur. C'est pour cela que plusieurs personnes engagées dans le développement de modèles en dehors comme au sein de l'administration ont souhaité à un moment ou à un autre accéder au modèle de l'Insee. En particulier, au lancement de l'Institut des politiques publiques (IPP), A. Bozio voulait récupérer la façon dont étaient codés les barèmes de l'impôt, plutôt que d'avoir à refaire l'ensemble du travail de reconstitution. Rendus en accès libre sous l'appellation « barèmes IPP », ces derniers s'inscrivent dans un mouvement général d'ouverture des codes amorcé à la fois de l'intérieur (3.3.1.) et de l'extérieur de l'administration (3.3.2.).

3.1. Les premiers pas d'une culture du « libre » au sein de l'administration

A la fin des années 2000, deux économistes du Centre d'Analyse Stratégique placés auprès du Premier Ministre (CAS, devenu « France Stratégie » depuis) souhaitent disposer d'un instrument de simulation de la législation socio-fiscale. Le CAS ne faisant pas partie du service statistique publique, ils n'ont pas accès aux versions les plus récentes de l'ERF, ni aux modèles

[34] Par exemple, les données de justice peuvent être consultées par les chercheurs, mais aucun traitement ne peut être réalisé.

de microsimulation déjà existants au sein de l'administration. Les deux économistes décident alors de développer leur propre programme, baptisé Openfisca[35]. Celui-ci permet initialement de calculer un montant d'impôt et de prestations sociales à partir d'une situation individuelle donnée et alimente bientôt un simulateur en ligne à destination des bénéficiaires. Il connaitra par la suite de nombreuses évolutions, en particulier suite à l'association – formelle puis informelle, d'Etalab et de l'IPP autour du modèle suite à une commande du cabinet du premier ministre de l'époque suivie par Fabien Dell. Partageant la culture du logiciel libre, les auteurs choisissent dès le début de mettre leur instrument en libre accès, inspirés par le regroupement des forces intervenu dans un autre domaine :

> « Auparavant en météorologie, tous les labos faisaient leur propre modèle de simulation de la météo du monde et s'épuisaient dans la course à la construction du modèle parfait. Un des labos qui a ouvert son modèle, et tout le monde s'est mis à collaborer sur le même modèle. […] Et je me suis dit que c'est exactement à ça qu'on est confronté aujourd'hui : plusieurs administrations qui construisaient leurs modèles, en concurrence les unes avec les autres. »[36]

Le code est mis en ligne sur une plateforme contributive (Github)[37], qui permet de suivre les améliorations proposées par un nombre potentiellement important de développeurs, et de les contrôler avant intégration. Le choix de passer par cette plateforme est aussi une façon pour ses créateurs d'empêcher un « verrouillage » ultérieur par l'administration[38].

En parallèle du recours à une communauté élargie de développeurs, les deux membres du CAS échangent très tôt avec les équipes de microsimulation de l'Insee et de la Drees, favorables à une mutualisation du travail de mise à jour annuelle de la législation socio-fiscale. Du côté de ces administrations chiffreuses traditionnelles, le renouvellement des équipes voit également le modèle Ines passer aux mains d'agents plus en phase avec la philosophie du logiciel libre. Schématiquement, nous pouvons distinguer dans les différentes équipes en charge du modèle d'Ines depuis sa création des profils plus orientés vers la programmation, d'autres profils plus orientés vers les études, sans que cette distinction ne recoupe initialement des différences de formation. Se succèdent au sein des mêmes équipes des personnes aux parcours comparables plus ou moins préoccupées par les approximations faites par Ines et plus ou moins disposées à produire des résultats ou affiner le modèle. La tendance à privilégier le travail sur le modèle plutôt que la production de résultats a pu être déplorée par l'une des enquêtés à partir de l'expression « geek de trop » :

> « Il y avait toujours ce que j'ai appelé le geek de trop : c'est la personne qui en voulant bien faire va écrire des pages et des pages de programme pour essayer de bien simuler quelque chose, mais qui va oublier que derrière le coût d'entrée pour les autres va augmenter et que finalement c'est aussi risque d'erreur ». (Entretien avec un membre de l'équipe Ines-Insee).

Ce qui apparaît au « geek » comme une amélioration indispensable, voire de bon sens, peut être perçu comme une source de complexité supplémentaire pour d'autres membres de l'équipe. Ce décalage apparaît dans la manière dont a pu être appréciée une nouvelle refonte du programme intervenue au début des années 2010 à la seule initiative de l'équipe Insee, qui a dû non seulement convaincre sa hiérarchie mais aussi passer outre les réserves de l'équipe Drees[39].

---

[35] Voir Shulz (2019) pour une analyse détaillée du développement de ce modèle.

[36] Témoignage d'un des deux auteurs du modèle cité par Shulz (2019), p 854.

[37] Voir le chapitre 4 pour une mise en perspective des plateformes digitales d'innovation.

[38] Entretien avec le second auteur du modèle, août 2020.

[39] Cette nouvelle version d'Ines est moins facile d'accès car elle rompt avec la logique linéaire qui prévalait jusqu'ici en multipliant les bouts de programme appelés à différents endroits d'un fichier maître. Cette structure apparaît évidente aux « geeks » suivants, qui auraient tendance à relativiser l'ampleur du changement, et peu

Ce changement est intervenu à une époque où les équipes de l'Insee et de la Drees se contentaient de transmettre leurs modifications de programme par mail, ce qui pouvait créer des décalages entre les versions manipulées de part et d'autre. Ces défauts de coordination ont été surmontés par un changement d'organisation rendu possible par un rattrapage technologique. A partir de 2014, les équipes de l'Insee et de la Drees ont eu recours à la plateforme en ligne Adullact[40] pour partager le modèle en conservant la trace des versions successives des programmes, de manière à pouvoir à la fois intégrer directement toutes les modifications de bouts de codes dans une version unique d'Ines et revenir facilement en arrière à tout moment. Il s'agissait là encore d'une évolution de bon sens d'un point de vue de développeur informatique puisque le recours à un logiciel de versionnage était déjà à cette époque une pratique courante en dehors de l'administration. L'initiative en est revenue à une administratrice Insee au profil atypique, puisque cette dernière a intégré l'Ensae alors qu'elle préparait un doctorat de physique théorique et mathématique, à une époque où les parcours « data science » de l'Ensae n'avaient pas encore été créés. Durant son passage à la Division des Études Sociales, elle a travaillé à l'optimisation d'Ines, si bien que ce qui tournait en deux heures a fini par tourner moins d'un quart d'heure[41], permettant encore de nouvelles générations d'études[42]. Son expérience a pu convaincre la direction de l'Insee de l'utilité de créer temporairement un poste supplémentaire destiné à un profil informatique pour poursuivre le développement du modèle.

Entre le début des années 2000 et le milieu de la décennie suivante, la figure du « geek de trop » laisse ainsi place à celle plus valorisée du « data scientist » qui est parvenu à démontrer l'intérêt de continuer à investir dans le code tout en fluidifiant la coordination entre les équipes, rendant par-là possibles de nouveaux rapprochements avec de nouvelles équipes de microsimulation, motivés d'une part par une conception élargie de la statistique publique, d'autre part par la philosophie du logiciel libre. Facilité par le passage sur Adullact, l'ouverture d'Ines est défendue par les agents en charge du modèle à partir d'arguments de différents ordres :

> « C'était vraiment une volonté de la part des agents tant du côté de l'Insee que de la Drees, de participer à ce mouvement vers le libre, de faire augmenter la connaissance globale, avec l'idée que cela pouvait nous servir aussi, parce des utilisateurs pourraient participer à mettre à jour ou améliorer le code. Pour être très honnête sur ce point, on n'a pas vu d'énormes gains pour nous, à part quand même un gain qui est très important, un gain de réputation : l'Ofce nous a fait beaucoup de pub par exemple en disant que cette ouverture était une innovation majeure pour la statistique au XXIe siècle. Et puis derrière cela aussi il y a le fait que c'est trop coûteux d'avoir autant de modèles de microsimulation, et donc on pensait qu'il allait probablement y avoir un moment où il y aurait un rassemblement, et donc qu'il valait mieux mettre Ines en libre le plus rapidement possible pour le faire davantage connaître » (Entretien avec Michael Sicsic, responsable du modèle Ines à Insee de 2015 à 2020, juillet 2020).

intuitive aux autres. Une évolution qui a enrichi les potentialités de l'outil, notamment en permettant d'appliquer les législations socio-fiscale de différentes années à une même base. Mais elle a aussi nécessité un temps important d'appropriation et de contrôle des conséquences de l'ensemble des changements.

[40] Association des Développeurs et Utilisateurs de Logiciels Libres pour les Administrations et les Collectivités Territoriales.

[41] Deux heures pour faire tourner l'ensemble du programme, actualisation de la base comprise. Ce résultat est aussi en partie dû à la mise en place des serveurs centraux à l'Insee.

[42] En particulier, le calcul de l'indicateur de pauvreté avancé qui s'est tout de suite imposé dans le paysage de la statistique publique.

Le début des réflexions sur l'ouverture d'Ines datent de 2012, avant même le passage sur Adullact. Le modèle a été ouvert en 2016[43]. Dans cet intervalle, les échanges avec les développeurs d'OpenFisca ont été brusquement interrompus par l'irruption d'autres militants du libre au mode d'action plus offensif. Ceux-ci vont pousser à une ouverture généralisée des modèles de microsimulation.

3.2. Ouverture forcée : l'administration sommée de communiquer les codes de ses modèles de microsimulation

En 2014, un jeune économiste en stage de fin d'études au Secrétariat Général de Modernisation de l'Action Publique (SGMAP), et plus précisément au sein du département Etalab[44] chargé de coordonner et mettre en œuvre la stratégie de l'Etat en matière d'ouverture des données des administrations, contribue au développement d'OpenFisca (un partenariat est à l'époque conclu entre ses concepteurs au sein du CAS, et Etalab). Il a alors pour tâche la transcription en code de la législation concernant des niches fiscales de l'impôt sur le revenu. Plutôt que de recoder entièrement cette législation, il envisage d'utiliser directement le calculateur utilisé pour recouvrir l'impôt[45]. Ses premières demandes directes refusées, il s'engage dès mai 2015 dans des procédures judiciaires. Ce dernier n'aura eu recours à aucun avocat, et a saisi la justice sans disposer de moyens particuliers. Il saisit ainsi lui-même la CADA, puis le tribunal administratif pour obtenir finalement le code en question en mai 2017, une semaine avant la délibération finale, soit 20 mois après le début de sa démarche. Entre temps, la Direction générale des finances publiques organise un hackathon en avril 2016 (Algan, Bacache-Beauvallet, Perrot, 2016). Réunissant près de 150 participants issus aussi bien de l'administration que de la société civile (entreprises, chercheurs, « simple » citoyen) et organisé avec Etalab, l'objet de ces deux journées était d'imaginer des usages possibles en dehors de l'administration… du code source du calculateur des impôts. La procédure judiciaire en cours n'est évoquée ni par le ministre des Finances et des Comptes publics de l'époque Michel Sapin, ni par le secrétaire d'État chargé du budget Christian Eckert, ni la secrétaire d'Etat chargé du numérique Axelle Lemaire (alors en charge de la Loi pour une République Numérique, présentée en conseil des ministres quelques jours plus tôt). Cette dernière souligne plutôt le fait que *« l'ouverture d'un code source d'un calculateur par une administration est une première (…) cette démarche (…) est l'image d'un État ouvert et responsable, qui considère que la gestion de notre « bien commun » n'est pas une prérogative de l'administration, mais que c'est une responsabilité de tous »*[46].

Quoi qu'il en soit, la démarche engagée auprès de la CADA fera jurisprudence[47], et va directement inspirer la création d'une association quelques mois plus tôt : l'Ouvre-boîte. Crée le 6 mars 2017, cette association a pour objet l'obtention de « l'accès à la publication effective des documents administratifs, et plus particulièrement des données, bases de données et codes sources, conformément aux textes en vigueur »[48]. Ses membres coordonnent leur activité via

[43] https://adullact.net/projects/ines-libre/

[44] Voir Goëta (2017) pour une présentation de cette structure créée en 2011.

[45] Ce dernier raconte, sous forme d'un tutoriel, pourquoi et comment il a entrepris de demander l'ouverture de ce code, sur le forum d'Etalab : https://forum.etalab.gouv.fr/t/howto-obtenir-dune-administration-lacces-a-un-code-source/186

[46] Discours d'introduction d'Axelle Lemaire, 1er avril 2016. https://www.economie.gouv.fr/files/files/PDF/discours_axellelemaire_hackathon.pdf

[47] Les modèles de microsimulation sont désormais considérés comme des documents administratifs, et peuvent donc être communiqués à ce titre aux personnes qui en feraient la demande.

[48] Annonce n°1903 p113 publiée au Journal Officiel portant création de l'association.

différentes plateformes web (wiki d'information, forum de discussion), et ont défini un certain nombre de règles afin d'optimiser leur coordination et limiter les risques d'erreurs[49].

> « Il a déroulé tout le processus sans aide extérieure, c'est lui qui a rédigé sa demande, son mémoire... L'Ouvre-boîte n'existait pas, et il a prouvé que c'était possible d'actionner les leviers juridiques sans avoir de gros moyens. On a repris la même méthode, qu'on a poursuivi dans la durée (…). Cet exemple nous a prouvé que les démarches nécessaires pour ouvrir des documents demandent une énergie limitée. » (Extrait entretien avec un membre de l'ouvre-boîte, juillet 2019)

Les codes sources étant désormais considérés comme des documents administratifs comme les autres, l'association entreprend des demandes pour obtenir (avec succès) l'ouverture des calculateurs de l'impôt, et obtient même l'ouverture complète de 3 années de calculateurs de l'IR et de l'ISF. Avant même la demande CADA, Etalab, qui a été créé par décret en 2011, était déjà connu pour pousser les administrations à ouvrir leurs données au nom de l'open-data. Des demandes avaient déjà été formulées en ce sens directement auprès des équipes de microsimulation ainsi qu'auprès de leur hiérarchie.

Mais à l'image du simulateur *Révolution Fiscale,* la demande d'ouverture du code des impôts via la CADA va crisper les administrations. Cette demande émanant d'une personne chargée de contribuer au développement d'Openfisca, les discussions entre les équipes de microsimulation au sein de l'Insee-Drees et Openfisca sont rompues. En outre, ces dernières sont critiques de la méthode du « couteau sous la gorge », pensant que la méthode était approuvée par le département Etalab. Etalab explique de son côté que la démarche a été réalisée à l'initiative de la personne en question, et que la politique interne voulue par le directeur de l'époque était de conserver de bonnes relations avec les autres administrations, en particulier avec Bercy.

> « Et puis après s'est rajouté l'histoire des demandes CADA, qui ont détérioré tout le dialogue qu'il y avait (…). La première demande d'accès s'est faite sur le code source des impôts je crois (…). Et bien je crois que c'était la dernière fois qu'on avait Etalab aux groupes de travail. Je pense que ça a crispé tout le monde qu'on discute d'un côté et que de l'autre côté ils mettent le couteau sous la gorge (...) : " vous ne voulez pas travailler avec nous, et bien on va vous y forcer en fait "» (Entretien avec un membre de l'équipe Ines, juillet 2020).

Il n'empêche que l'association Ouvre-boîte sera fondée quelques mois seulement après cette demande CADA par, entre autres, le superviseur du stagiaire en question. L'association va alors entreprendre plusieurs dizaines de demandes d'ouvertures de codes, sur une multitude de sujets. Parmi elles, des demandes d'ouvertures d'un des modèles de microsimulation dynamique de l'Insee, Destinie, quelques mois seulement avant la mise en *open-access* sur la plateforme Adullact du modèle de microsimulation statique de l'Insee-Drees.

Désignation du document	**Administration détentrice**	**Date de demande**	**Date ouverture**
Bilan GES	ADEME	02/03/2020	11/06/2021
Inventaire du patrimoine naturel	OFB-CNRS	05/01/2021	05/01/2021
Comptes	Sciences po	18/11/2018	18/03/2021

[49] https://ouvre-boite.org.

Données de trafic SURF3	Ville de Paris	28/12/2018	25/07/2020
Base Carbone	ADEME	02/03/2020	28/05/2020
Données du Budget	Direction du Budget	09/05/2018	12/05/2019
Bases de l'Inpi	Inpi	29/05/20217	17/04/2019
Modèle Mélèze	Insee	15/11/2017	26/03/2019
Cartes géologiques "Bd Charm-50"	Brgm	08/04/2018	07/03/2019
Modèle Omphale2010	Insee	15/11/2017	21/12/2018
Modèle Myriade*	Cnaf	15/11/2017	27/09/2019
Modèle Destinie2	Insee	08/11/2017	20/09/2018
Modèle Saphir	DG Trésor	08/11/2017	05/09/2018
Modèle Mésange	DG Trésor	08/11/2017	05/09/2018
Catalogue des collections des musées de France Joconde (extrait)	Musées de France	14/05/2017	03/04/2018
Cadastre	DGFiP	27/06/2017	29/09/2017
Calculette de l'impôt sur le revenu	DGFiP	04/05/2017	14/09/2017

** Le modèle Myriade n'a pas fait l'objet d'un recours à la CADA (un premier recours avait été adressé au ministère de la santé, qui n'a pas la main sur le modèle.*

Tableau 2. Les demandes formulées par l'association Ouvre-boîte depuis sa création et ayant abouti, classées par date décroissante d'acceptation (Source : site internet de l'association).

Les demandes d'ouverture des modèles sont motivées par une raison première, qui relève, pour les membres de l'association, de ce que l'on pourrait désigner comme une forme de statactivisme démocratique : pouvoir librement auditer l'administration et réutiliser les modèles qu'elle développe au moyen de deniers publics, pour d'autres usages afin de contribuer au débat public (parti politique, journalisme, décodeurs, cabinets d'économistes, laboratoires de recherche…).

Le mode d'intervention axé sur l'affrontement direct avec l'administration est clairement assumé, pour un souci d'efficience : selon ses membres, l'administration ne se prêtant pas toujours au jeu, cela ne vaut pas la peine de consommer du temps pour par exemple l'assister dans une éventuelle démarche d'ouverture :

> « Avec nos moyens très limités, nous ne faisons que le strict nécessaire. Ce qui nous rend bourru aux yeux des administrations. De plus, il est difficile de tisser une relation de confiance et dans le même temps d'intenter un procès. D'autres acteurs de la société civile tentent d'instaurer un dialogue, mais se retrouvent limités dans leurs moyens d'action. L'orientation juridique de l'association nous libère de ces remords, mais nous enchaîne souvent dans un rapport de confrontation » (Entretien avec un membre de l'Ouvre-boîte, juillet 2019).

Ce positionnement permet d'offrir à certains acteurs qui ne voudraient pas entacher de bonnes relations avec l'administration, un cadre pour forcer l'ouverture sans avoir à s'exposer publiquement – les démarches pouvant être effectuées et signées par un autre membre de l'association. Les membres de l'association se font d'ailleurs plutôt discrets, en dehors de quelques anciens membres d'Etalab faisant partie des membres originels de l'association. Cela s'explique par les convictions personnelles de ces anciens agents, et non par une volonté d'Etalab de disposer d'un organisme permettant de formuler des demandes CADA (pour rappel, Etalab ne pourrait pas formuler lui-même des demandes à la CADA, cette dernière ne statuant pas sur les relations inter-administrations ; les recours devant par ailleurs être effectués par une personne physique). De plus, l'association considère que ces demandes constituent une forme d'acculturation aux administrations. En s'appuyant sur la loi LPR de 2016 qui permet d'ouvrir par défaut de nombreuses données et codes, le pari est pris que les demandes effectuées auprès des administrations permettront de leur faire prendre conscience que « ce n'est pas si compliqué d'appliquer l'open-data par défaut » (extrait d'entretien avec un représentant de l'Ouvre-boîte, juillet 2019).

Quelles conséquences ont eu ces demandes sur la façon d'ouvrir les modèles, au sein de la Direction générale du Trésor (modèle Saphir) et de la Cnaf (modèle Myriade) ? Côté Cnaf, le modèle ayant été abandonné depuis plusieurs mois, l'ouverture ne présentait pas un enjeu pour l'administration. Les seules réticences sont venues du sentiment « de perdre du temps ». Côté DG Trésor, une partie des agents étaient en faveur de cette ouverture pour offrir plus d'« accountability » et de transparence aux citoyens. Une autre partie était réticente, anticipant des « problèmes » auxquels la direction serait confrontée en cas d'ouverture (que l'on retrouvait déjà au moment des discussions sur l'ouverture du modèle de l'Insee-Drees) : contre-chiffrage à produire pour démentir une mauvaise utilisation du logiciel, vérification de l'ensemble des travaux antérieurs du Trésor (et donc, détection de potentielles erreurs, qui entacheraient la crédibilité de l'administration), attaques de la part de la Presse etc. Finalement la solution retenue dans un premier temps par le Trésor, aussi bien pour son modèle de microsimulation Saphir que pour les autres modèles pour lesquelles des demandes ont été formulées (Mésange et Opale) par l'Ouvre-boîte, a été de s'en tenir à une stricte conformité au droit, sans profiter de l'occasion pour adopter une démarche plus volontaire comparable à celle de l'Insee. Toutefois, plus récemment, le Trésor semble avoir évolué sur cette question. Depuis juin 2021, leur site met à disposition des programmes actualisés du modèle macroéconomique Opale et

communique plus volontiers, via les réseaux sociaux sur la mise à disposition de leur modèle susceptible de leur apporter quelques bénéfices en termes d'image50

Figure 2. La directrice de la DG Trésor twitte sur la mise en ligne d'un complément d'un de ses modèles. Un chercheur de l'Ofce réagit (juin 2021)

Cette politique volontariste de la DG Trésor peut d'une part se lire comme le signe que la stratégie consistant à forcer l'ouverture pour changer la culture des administrations serait fructueuse, puisque c'est après une première série d'ouvertures « forcées » que la DG Trésor s'est engagée dans la mise en ligne de ces programmes actualisés. D'autre part, cela pourrait également être vu comme le suivi de la voie montrée par l'équipe Ines en ouvrant son modèle de façon spontanée ; ou bien comme le fait que l'enjeu se déplace désormais de l'ouverture à la valorisation des données ouvertes. S'il n'est pas possible de trancher sur l'une ou l'autre des lectures, nous souhaitons à présent enrichir cette réflexion d'un regard plus global sur les résultats obtenus par les différentes façons d'obtenir des codes et de la donnée.

---

[50] « Tresthor : le nouvel outil de la DG Trésor pour réaliser des prévisions macroéconomiques », https://www.tresor.economie.gouv.fr/Articles/2021/06/30/tresthor-le-nouvel-outil-de-la-dg-tresor-pour-realiser-des-previsions-macroeconomiques. Le ton de cet article publié par la DG Trésor le 30 juin 2021, assorti d'une communication sur les réseaux sociaux réalisé notamment par son économiste en chef, contraste avec la communication minimale effectuée 3 ans plus tôt pour les ouvertures forcées par l'Ouvre-boîte : https://www.tresor.economie.gouv.fr/Articles/2018/09/05/la-dg-tresor-met-a-la-disposition-du-public-les-codes-sources-des-modeles-mesange-opale-et-saphir

4. Discussion : différentes conceptions de la qualité de l'ouverture

Pour finir, nous souhaitons interroger les différentes exigences en matière d'ouverture portées par les principaux protagonistes de cette histoire et comparer à partir des résultats obtenus à ce jour l'intérêt, pour de potentiels utilisateurs de modèles de microsimulation, des différentes démarches entreprises. Plus généralement, l'étude de ces deux dimensions permet de comprendre la coexistence de plusieurs conceptions de la qualité de l'ouverture, c'est-à-dire de différentes manières de définir ce qu'est une bonne ouverture.

Cette enquête nous a d'abord permis de mettre au jour une tension entre deux attentes vis-à-vis de la mise à disposition des codes des modèles de microsimulation de l'administration : la transparence et l'accessibilité. Du point de vue de l'accessibilité, « bien ouvrir » consiste à rendre le code le plus directement utilisable. Cette conception qui semble la plus courante est mise en avant par les auteurs d'un des tous premiers guides de l'open-data paru en France (Chignard et Marchandise, 2012). Elle a également motivé un investissement consenti par les membres de l'équipe Ines au moment d'ouvrir leur modèle[51]. En plus de rédiger une notice de présentation de leur instrument de microsimulation, ces partisans du modèle libre ont dû modifier leurs programmes pour répondre à cette exigence d'accessibilité. D'abord, ils ont ajouté des commentaires pour rendre le code plus clair pour des utilisateurs extérieurs à l'Insee et à la Drees. Ensuite, ils ont cessé d'utiliser certaines variables de l'ERF auxquelles les utilisateurs extérieurs au Service Statistique Public ne pouvaient pas avoir accès (y compris via le Réseau Quetelet), comme la commune ou d'autres informations « indirectement nominatives » susceptibles de permettre dans certains cas d'identifier par recoupement une personne particulière. Contraintes d'ouvrir leur modèle par des demandes extérieures, d'autres équipes de microsimulation comme celle de Myriade ou Saphir, se sont au contraire contentées de mettre à disposition le code existant sans se soucier des possibilités d'utilisations ultérieures, et en y consacrant le moins de temps possible[52]. Cette seconde manière d'ouvrir est plus conforme aux attentes de l'association l'Ouvre-boîte, qui se concentre avant tout sur la stricte application du droit. Guidés par une exigence de transparence, ses membres considèrent plutôt que bien ouvrir consiste avant tout à livrer sans restriction le code brut, tel qu'utilisé en l'état. Dans ces conditions, il n'est pas attendu de l'administration qu'elle fournisse une notice explicative ou « nettoie » son code, ce qui peut d'ailleurs être interprété comme une façon de cacher des imperfections. De fait, l'une des craintes soulevées par le projet d'ouverture d'Ines était que les équipes se trouvent submergées de demandes de justifications par le grand public. Et le principal effort consenti par la Cnaf au moment de l'ouverture forcée de Myriade a été de relire l'ensemble du programme et de vérifier qu'il n'y avait pas dans les commentaires d'éléments susceptibles d'être contestés.

Au-delà de l'absence de retouche des documents transmis, la conception de ce que doit être une bonne ouverture guidée par l'exigence de transparence suppose l'absence de toute barrière à leur consultation. A ce titre, les conditions d'accès au modèle Ines ont également fait l'objet de critiques. Pour consulter le programme, les utilisateurs extérieurs doivent créer un compte sur la plateforme Adullact puis demander à rejoindre le projet « Ines Libre », ce qui limite

[51] L'exigence d'accessibilité commanderait même que les équipes abandonnent le logiciel propriétaire sur lequel tourne leur modèle (SAS) pour garantir une réelle ouverture. C'est bien ce qu'elles sont en train de faire (en passant sous R) mais pour d'autres considérations (budgétaires), indépendantes du projet d'ouverture.

[52] Contrairement à Ines (Albouy et al., 2003 et les encadrés réguliers dans les publications produites à partir du modèle) et Myriade (Legendre, Lorgnet, Thibault, 2001 et Marc et Pucci, 2011), Saphir n'avait jusqu'à son ouverture faire l'objet d'aucune publication le présentant. Par conséquent, la DG Trésor a quand même rédigé une note de présentation du modèle au moment de l'ouvrir (Amoureux, Benoteau et Naouas, 2018), là encore en y consacrant moins d'attention que pour Ines.

doublement l'ouverture du modèle aux yeux d'une partie des membres de l'association l'Ouvre-boîte :

> « On peut y avoir accès au bout d'un temps, qui est indéterminé puisqu'il y a une validation manuelle. Si la personne qui valide est en vacances, ça peut prendre du temps. Ce n'est pas conforme aux textes, et ça irrite ceux qui sont habitués aux usages des communautés open-source. Autre problème si l'accès n'est pas anonyme : c'est une restriction qui n'est pas non plus prévue par les textes et qui va à l'encontre des pratiques open-source » (Extrait d'entretien avec un représentant de l'Ouvre-boîte, juillet 2019).

Du point de vue des équipes de microsimulation de l'Insee et de la Drees, l'un des objectifs de l'ouverture d'Ines était de créer une communauté d'utilisateurs susceptible d'apporter des améliorations au modèle. L'existence de cette communauté peut être facilitée par la connaissance des personnes qui accèdent au modèle et ces dernières sont d'ailleurs pour la plupart bien identifiées par un responsable de l'équipe d'Ines côté Insee que nous avons pu interroger. La procédure d'inscription préserve toutefois la possibilité pour les utilisateurs de rester anonymes.

Si l'on se penche maintenant sur les effets produits par les différentes demandes d'ouverture, on distingue deux stratégies (forcer l'ouverture / gagner la confiance de l'administration) et deux cibles prioritaires (les codes / les données). En matière de microsimulation, la stratégie de l'Ouvre-boîte s'est concentrée sur les codes dont elle a cherché à forcer l'ouverture. La mise en ligne du modèle de microsimulation encore utilisé aujourd'hui par la DG Trésor constitue sans doute son résultat le plus spectaculaire. En dehors des agents de cette direction, rares étaient les personnes à savoir à quoi ressemblait l'architecture du modèle et la façon dont un dispositif particulier était précisément chiffré par le Trésor. Une non-transparence qui pouvait s'expliquer par le fait que, à la différence de l'Insee ou la Drees, la DG Trésor réalise la plupart de ses travaux en réponse à des commandes émanant du cabinet du ministre[53].

Désormais, tout le monde peut donc télécharger d'un simple clic sur le site internet du Trésor les 15 programmes qui constituent le modèle et consulter par exemple la manière dont est simulé le non-recours à la prime d'activité et au RSA (programme 14), directement sous SAS (le logiciel prioritaire sous lequel tourne le modèle) mais aussi à partir de n'importe quel lecteur de fichier texte. L'intérêt de pouvoir ainsi lire le programme apparaît néanmoins limité aux yeux d'acteurs de la microsimulation extérieurs à l'administration que nous avons pu interroger, qu'il s'agisse d'un membre de l'IPP qui développe son propre modèle ou même d'un membre de l'OFCE qui s'était pourtant félicité au moment de l'ouverture d'Ines, de pouvoir récupérer un modèle de l'administration[54] :

> « Contrairement à Saphir, Ines n'est pas ouvert de force, il est mis à disposition, ce qui fait une grande différence. Au mois de novembre quand l'équipe Ines fournit la législation de l'année N-1, ça tourne : vous avez le fichier SAS, vous appuyez sur « Entrée », vous avez les bonnes tables de l'ERF, et le modèle (…) calcule tous les agrégats. Tandis que pour Saphir, vous n'avez que du code. [Or] pour évaluer un code il faut pouvoir l'exécuter, et [donc] ce qui est essentiel c'est d'avoir en entrée les données dans le format utilisé par le code. »

[53] La Direction générale du Trésor réalise toutefois des travaux à une fin d'étude et de recherche, notamment pour la revue académique *Économie & Prévision* qu'elle édite. Elle a également intensifié sa communication extérieure plus récemment à travers ses notes thématiques *Trésor-Eco*.

[54] Madec Pierre et Timbeau Xavier, « Statistique publique : une révolution silencieuse », OFCE le blog, 2017.

D'autres raisons sont également mises en avant par Antoine Bozio pour relativiser l'importance des codes, tout en durcissant ce clivage entre intérêt pour le code et intérêt pour les données.

> « Je ne vois pas l'ouverture du code de Saphir obtenu par l'Ouvre-boîte comme quelque chose d'essentiel (…). Pour moi, l'information importante, ce n'est pas le code, ce sont les données sous-jacentes (...). Dans la microsimulation, il y a ceux qui croient au code et ceux qui croient aux données. Ceux qui croient au code pensent que c'est en faisant des cas-types qu'on comprend ce qui se passe (…). Mais c'est trompeur parce qu'on ne sait pas où se trouvent la masse des ménages ou des individus dans la distribution. Vous pouvez ainsi regarder des cas-types qui n'existent pas ! (…) Ce qu'il faut absolument savoir, ce sont les caractéristiques des ménages pour bien comprendre combien sont touchés par telle ou telle mesure. Et pour ça, il nous faut des données détaillées sur l'ensemble des ménages (...). Si on compare Ines et Taxipp à partir de leur code, on ne va pas trouver de différences substantielles, tandis que sur les données sous-jacentes l'information disponible pourra être réellement différente. » (Entretien avec Antoine Bozio, juillet 2021).

Nous l'avons vu, cette vision n'est pas seulement celle d'un microsimulateur externe à l'administration, mais aussi celle de l'un des principaux entrepreneurs du dégel des relations entre les chercheurs et les membres du SSP, pour l'accès aux données des seconds par les premiers. Son action traduit une autre forme de statactivisme, autant éloignée de celle de ses prédécesseurs, qui ont privilégié le contournement de certaines règles (ne pas réutiliser pour d'autres projets les données obtenues pour un projet spécifique), que de celle des datactivistes qui privilégient l'affrontement en s'appuyant sur d'autres règles (saisir la CADA puis le tribunal administratif pour ouvrir les données). Cette autre stratégie fait le pari d'un croisement de celle de la « négociation-discussion » décrite par les néo-institutionnalistes sociologiques, et de « l'implémentation-appropriation » (Pressman et Wildasky, 1973) de ses valeurs par une série d'interactions de proximité. Elle repose ainsi sur l'idée selon laquelle, « pour que cela marche, il ne faut pas s'antagoniser les administrations, mais gagner leur confiance » car « l'administration a tous les moyens de ne pas faire ». Notons que cette posture ne peut être ramenée simplement à une question de tempérament. Elle requiert au contraire des ressources relationnelles (la proximité avec des membres de cabinets qui eux-mêmes disposent d'une légitimité à agir sur ces questions, en l'espèce via l'instrument législatif) et institutionnelles spécifiques (A. Bozio est un universitaire à la tête d'un institut de recherche d'une grande école parisienne qui se donne précisément comme ambition d'être un des principaux acteurs de la microsimulation). Ces ressources ne sont pas les mêmes que celles à disposition des membres de l'association l'Ouvre-boîte, quand bien même certains sont membres de l'administration. Pour ces derniers, le répertoire d'action mobilisable apparaît ainsi plus naturellement celui de l'usage du cadre légal et sa dimension coercitive ; d'autant que passer par une structure associative leur permet de ne pas agir en leur nom propre, et d'ainsi préserver leur position au sein de l'administration.

Dans la logique moins coopérative qui prévaut pour les deux autres stratégies, les parties contournées – ou attaquées « à l'ouvre-boîte », peuvent aussi mobiliser d'autres ressources juridiques pour se défendre et contrarier le mouvement d'ouverture. Le Trésor a ainsi pu se retrancher derrière l'argument du secret des délibérations du gouvernement pour ne livrer qu'une partie de ses programmes (le modèle de base) et conserver les codes développés dans Saphir pour simuler différents projets de réformes. Sommés à leur tour par l'association Ouvre-boîte de publier leurs codes, les membres de l'Ofce ont quant à eux invoqué, avec succès, le droit de la propriété intellectuelle des chercheurs pour maintenir fermés leurs modèles. Une parade qui permet d'éviter que les démarches de l'association soulèvent de nouvelles questions juridiques relatives à l'accès aux résultats de partenariats de recherche public-privé. Par ce retournement inattendu, les chercheurs, qui semblaient a priori pouvoir le plus bénéficier des

combats des militants de l'ouverture, sont donc devenus à la fois les nouvelles cibles et l'un des principaux remparts à une généralisation de ce mouvement.

Enfin, et dans une perspective plus générale, l'histoire de l'ouverture des modèles de microsimulation et de leurs données permet de poser dans d'autres termes la question de la rationalisation des ressources – et des efforts mobilisés pour apprécier l'action publique à l'aune des outils numériques ; mais aussi des frontières de l'expertise de l'État en matière d'évaluation. D'un côté, les administrations semblent progressivement engagées vers un regroupement autour d'un seul et unique modèle. La mise à jour régulière du modèle nécessite en effet la mobilisation de plusieurs agents, et le maintien de différents modèles au sein des administrations – qui a pu présenter un intérêt certain pour identifier les premières erreurs au début de leur utilisation, semble de moins en moins justifiée. Et cela d'autant plus dans un contexte de contrainte budgétaire et de recherche continue d'optimisation des ressources existantes, déjà prégnant au moment des premiers mouvements d'ouverture (Pénissat, 2009). D'un autre côté, l'open data offre à des acteurs en dehors des seules administrations chiffreuses « historiques » la possibilité d'apporter une expertise évaluative d'une robustesse comparable à celle produite par ces dernières, pour peu qu'ils aient accès aux ressources nécessaires. S'ils ne produisent pas leur propre modèle, ces acteurs peuvent même directement participer à la coproduction avec l'administration du modèle via les plateformes opensource de gestion de versions. De fait, la question de la rationalisation se pose à la fois au sein des administrations, mais aussi en dehors, puisqu'ils disposent désormais non pas d'un mais de plusieurs modèles développés et mis à jour.

Finalement, bien ouvrir n'est-ce donc pas aussi savoir fermer ? Si la réponse à cette question semble largement affirmative pour la plupart des personnes interrogées, se pose néanmoins la question de quels modèles fermer. Du point de vue de l'utilisateur du modèle libre de l'administration (Ines), ce dernier peut être vu comme le modèle « à conserver » pour des raisons historiques et de légitimité dans la certification conforme du code. Du point de vue de l'utilisateur d'un modèle libre développé en dehors du service de statistique public (Openfisca, utilisé par la dernière version de Taxipp), les arguments relèvent plutôt de la qualité technique des simulation (qualité du code et des données utilisés), en plus d'une réutilisation simplifiée et d'une communauté d'utilisateurs plus importante.

Sans prendre parti pour une ou l'autre des positions (chacune des communautés anticipant à moyen terme l'abandon du modèle d'en face), notons toutefois que l'Assemblée nationale a récemment choisi de mettre dans les mains des parlementaires le microsimulateur LexImpact[55], qui développé à partir d'Openfisca. Un signe que cet instrument est désormais plus qu'une « expérimentation institutionnelle menée par des hackers politiques aux marges de l'administration » (Schulz, 2019). Une voie intermédiaire se dessine ainsi : l'utilisation d'une base commune et développée par tous (la législation socio-fiscale) ; avec une mise en pratique différenciée en fonction des données accessibles et des usages recherchés, qu'il s'agisse de chiffrer un amendement, tester une réforme, ou évaluer l'action du gouvernement.

[55] En janvier 2022 le modèle, toujours en développement, permet de micro-simuler une réforme de l'impôt sur le revenu, les dotations aux communes, les cotisations et prestations sociales, et la CSG. Il se distingue des autres modèles en proposant de réaliser les évaluations à partir d'une interface qui propose à l'utilisateur de modifier directement les articles des lois concernés.

Bibliographie

Albouy, V., Bouton, F., Le Minez, S., Pucci, M. (2003). Le modèle de microsimulation Ines : un outil d'analyse des politiques socio-fiscales. *Dossiers Solidarité et Santé*, 3, 23–43.

Algan, Y., Bacache-Beauvallet, M., Perrot A. Administration numérique. *Notes du conseil d'analyse économique*, 34(7), 1-12.

Amoureux, V., Benoteau, I., Naouas, A. (2018). *Le modèle de microsimulation Saphir*, Documents de travail de la DG Trésor.

Blanchet, D., Hagneré, C., Legendre, F., Thibault, F. (2015). Microsimulations statique et dynamique appliquées aux politiques fiscales et sociales : modèles et méthodes. *Économie et Statistique*, 481(1), 5–30.

Béranger, J. (2017). *Quelle éthique pour une approche ouverte et communautaire de l'utilisation des big data en santé ?* [En ligne]. Disponible à l'adresse : https://medium.com/epidemium/quelle-%C3%A9thique-pour-une-approche-ouverte-et-communautaire-de-lutilisation-des-big-data-en-sant%C3%A9-e9c026881961 [Consulté le 27 octobre 2021].

Bessis, F., Cotton, P. (2021). La réforme, le chiffrage, son modèle et ses données. Les évolutions du monopole de l'expertise économique au prisme d'un instrument de microsimulation de la législation socio-fiscale, *Politix*, 134(2), 7–32.

Bozio, A. (2018). *Économie publique de la protection sociale*, mémoire pour l'habilitation à diriger des recherches, Aix-Marseille Université.

Bozio, A., Geoffard P.-Y. (2017). L'accès des chercheurs aux données administratives, Conseil nationale de l'information statistique, Paris.

Bruno, I., Didier, E., Prévieux, J. (dir.) (2014). *Statactivisme : comment lutter avec des nombres*, Paris, La Découverte, Zones, Paris.

Caporali, A., Morisset, A., Legleye, S., Richou, C. (2015). La mise à disposition des enquêtes quantitatives en sciences sociales : l'exemple de l'Ined. *Population*, 70(3), 567–597.

Chenu, A. (2011). Introduction. Dans *La France dans les comparaisons internationales. guide d'accès aux grandes enquêtes statistiques en sciences sociales,* Chenu, A., Lesnard, L. (dir.), Presses de Sciences Po, Paris., 9–17.

Chignard, S., Marchandise, J.-F. (2012). *L'Open data : Comprendre l'ouverture des données publiques*. FYP éditions, Paris.

Cobb, R. W., Elder C. D. (1972). *Participation in American politics. The dynamics of agenda-building*. Johns Hopkins University Press, Baltimore.

Concialdi, P. (2014). Le BIP40 : Alerte sur la pauvreté ! Dans *Statactivisme: comment lutter avec des nombres*, Bruno, I., Didier E., Prévieux J. (dir.), La Découverte, Zones, Paris, 199–211.

Goëta, S. (2015). Un air de famille : les trajectoires parallèles de l'open data et du big data. *Informations sociales*, 191(5), 26–34.

Goëta, S. (2017). Une petite histoire d'Etalab : comment l'open data s'est institutionnalisé en France ? *Statistique et Société*, 5(3), 11–17.

Landais, C. (2007). Les hauts revenus en France (1998-2006) : Une explosion des inégalités ? *Paris School of Economics Working Paper*.

Landais, C., Piketty, T., Saez E. (2011). *Pour une révolution fiscale : un impôt sur le revenu pour le XXIe siècle*, Seuil, La République des idées, Paris.

Legendre, F. (2019). L'émergence et la consolidation des méthodes de microsimulation en France. Économie et Statistique, 510-511-512, 201–217.

Legendre, F., Lorgnet, J.-P., Thibault, F. (2001). Myriade : le modèle de microsimulation de la CNAF. Un outil d'évaluation des politiques sociales. *Revue des politiques sociales et familiales*, 66(1), 33–50.

Marc, C., Pucci M. (2011). Une nouvelle version du modèle de microsimulation Myriade : trimestrialisation des ressources et évaluation du revenu de Solidarité active. *Dossiers d'études de la Cnaf*, 137.

Pénissat, E. (2009). L'État des chiffres : sociologie du service de statistique et des statisticiens du ministère du Travail et de l'Emploi (1945-2008). *Thèse de doctorat en sociologie, Paris, EHESS*.

Piketty, T. (1999). Les hauts revenus face aux modifications des taux marginaux supérieurs de l'impôt sur le revenu en France, 1970-1996. *Économie & prévision*, 138(2), 25–60.

Rhein, C. (2002). Démogéographie et données statistiques. *Espace Populations Sociétés*, 20(1), 125–132.

Shulz, S. (2019). Un logiciel libre pour lutter contre l'opacité du système sociofiscal. *Revue française de science politique*, 69(5), 845–868.

Silberman, R. (2011). *La protection des données individuelles en France et la recherche en sciences sociales*, Presses de Sciences Po, Paris.

Sujobert, B. (2014). Comment intervenir sur le programme de la statistique publique ? L'exemple des inégalités sociales. Dans *Statactivisme: comment lutter avec des nombres*, Bruno, I., Didier E., Prévieux J. (dir.), La Découverte, Zones, Paris, pp.213–231.

Pressman, J. L., Wildavsky A. B. (1973). *Implementation*, University of California Press, Berkeley.