ELIPSS : un dispositif longitudinal représentatif pour la recherche en sciences sociales

Stéphane Legleye[1,2,3], Anne Cornilleau[4], Anne-Sophie Cousteaux[4], Nirintsoa Razakamanana[1] et toute l'équipe CDSP et INED de coordination et de gestion du panel

Thématique : Enquêtes longitudinales

1. Institut national des études démographiques (INED), Paris (France)
2. Inserm, U669, Paris (France)
3. Univ Paris-Sud and Univ Paris Descartes, UMR-S0669, Paris (France)
4. Sciences Po, Centre de données socio-politiques, UMS828Paris (France)

L'équipe est composée de : CDSP : Emmanuelle Duwez, Adrien Ferreira, Daniele Guido, Laurent Lesnard, Matthieu Olivier, Geneviève Michaud ; INED : Gabrielle Bouchet, Carmen Calandra, Bernard de Clédat, Elodie Pétorin

Résumé

Internet est très utilisé par les sondeurs pour produire des enquêtes rapides et à bas coût. Pour cela, les instituts ont mis en place de larges panels d'internautes volontaires, qui peuvent s'y inscrire par exemple par des bannières postées sur des sites populaires. Ces panels sont donc constitués sans base de sondage ni procédure probabiliste, alors même que plus de 20% des ménages n'ont pas accès à Internet en 2013 [1].

ELIPSS (enquête longitudinale par Internet pour les sciences sociales) est un panel d'individus représentatif de la population des ménages ordinaires tiré aléatoirement dans le recensement, que l'on équipe gratuitement de tablettes tactiles connectées à Internet sans fil, en échange d'une participation mensuelle à 30 minutes de questionnaire. Ce panel fait partie d'un Equipement d'excellence *Données, infrastructures et méthodes en sciences humaines et sociales* (DIME-SHS) financé par l'ANR en 2010. ELIPSS est actuellement dans sa phase pilote (n=1000) et devrait atteindre la taille de 5000 individus fin 2015.

Dispositif d'enquête longitudinal original ouvert à la recherche, ELIPSS reste exposé à des biais de recrutement ainsi qu'à des risques d'attrition et de professionnalisation de ses membres qui pourraient en altérer la représentativité initiale. Il s'agit ici de présenter rapidement le panel et le recrutement du pilote, la structure sociodémographique de l'échantillon de répondants pour insister sur la présentation de l'évolution de la participation au cours du temps, de la typologie des trajectoires de réponse de ses membres aux enquêtes administrées depuis 2013 et enfin d'exposer le choix actuel du redressement.

Mots clefs : ELIPSS ; représentativité ; panel ; Internet.

Abstract

Survey agencies have developed large online panels offering the opportunity to conduct surveys at reduced costs. These panels are not probability-based: there is no sampling frame and potential members are volunteers recruited by invitations on popular websites. Additionally, as more than 20% of the households have no Internet access in France in 2013, they also suffer from a coverage defect.

ELIPSS (*enquête longitudinale par Internet pour les sciences sociales* - Internet longitudinal study for the social sciences) is a representative panel of individuals living in ordinary households that are randomly drawn in the national Census. Members are freely equiped with a digital tablet and a 3G internet connection in exchange of a monthly participation to a 30 minutes questionnaire. ELISS is part of the research infrastructure *Data, Infrastructure, Methods of Investigation in the Social Sciences and Humanities* (DIME-SHS) funded by the national research agency in 2010. It is currently in the pilote phase (n=1000) and should reach its final size of 5000 individuals in 2015.

Resource devoted to the research in social sciences, ELIPSS may suffer from recruitment bias and attrition, as well as professionalism of some of its members that may alter its initial representativeness. In this work, we first briefly expose the pilot panel and its recruitment and focus on the sociodemographic structure of the respondents, the evolution of the participation rate and the typology of the trajectories of individual participations to the first 9 achieved surveys. We finally present the current weighting procedure and discuss about the future challenges of the panel.

Keywords: ELIPSS; panel; representativeness; Internet.

Présentation du projet

Le projet a vu le jour pour permettre aux chercheurs de mener des enquêtes probabilistes sur des thèmes qui ne sont que peu traités par la statistique publique française et d'offrir des facilités pour conduire des enquêtes longitudinales de grande ampleur, ordinairement hors d'atteinte de la recherche académique en sciences humaines et sociales. ELIPSS doit permettre de diminuer les coûts et le temps de collecte et de faciliter l'administration et le suivi des enquêtes.

ELIPSS fait partie d'un Equipement d'excellence Donnés, infrastructures et méthodes en sciences humaines et sociales (DIME-SHS) financé par l'ANR en 2010. Cet equipex piloté par le Centre de données sociopolitiques (CDSP) de Sciences Po comprendégalement l'Ined (Institut national d'études démographiques), le Groupe des écoles nationales d'économie et statistiques (GENES), l'Université Paris-Descartes (UP-5), Telecom Paris Tech, le département recherche et développement d'Electricité de France (EDF R&D), l'Ecole des hautes études en sciences sociales (EHESS).

ELIPSS est un projet unique en son genre de panel à partir d'un échantillon probabiliste (représentatif) dont les membres sont équipés d'une tablette tactile connectée à Internet, en échange de leur participation à une interrogation mensuelle de 30 minutes au maximum. Ce dispositif est entièrement dédié à la recherche : les enquêtes sont proposées par des équipes de chercheurs des sciences humaines et sociales et sélectionnées par un comité scientifique et technique sur leur qualité méthodologique et scientifique.

Si ELIPSS est unique en son genre en raison de l'utilisation de tablettes connectées, il n'est pas pionnier dans le domaine des panels Internet représentatifs en population générale. En effet, des dispositifs similaires utilisant les ordinateurs personnels au sein des ménages ordinaires existent aux Pays-Bas (LISS Panel : http://www.lissdata.nl/ [2]) aux USA (Knowledge Panel : http://www.gfk.com/us/Solutions/consumer-panels/Pages/GfK-KnowledgePanel.aspx) et en Allemagne (GIP http://reforms.uni-mannheim.de/internet_panel/home [3], GESIS Panel : http://www.gesis.org/en/services/data-collection/gesis-panel/).

Le pilote ELIPSS

- La première phase du panel est sa phase pilote qui s'achèvera fin 2014. Le recrutement des panélistes y a débuté en 2012 et s'est achevé début 2013. Elle avait pour ambition de tester les procédures d'échantillonnage, de contact et de recrutement des panélistes, et définir les procédures de gestion du panel et de production des enquêtes et de développer les outils nécessaires. Il s'agit également de produire des estimations de l'attrition du panel pour préparer le plus précisément possible le cas échéant le passage à un panel à 5000 membres. Echantillonnage, critères d'éligibilité

Le panel ELIPSS est un panel aléatoire d'individus résidant en ménages ordinaires au sens de l'INSEE. Sont exclus les personnes sans domicile, les mineurs et les plus de 75 ans, les personnes résidant dans des collectivités (prisons, maisons de retraite, résidences étudiantes, etc.), mais aussi celles dans des logements précaires non considérés comme des logements au sens de l'INSEE et celles ne maîtrisant pas suffisamment la langue française pour répondre à des questionnaires autoadministrés.

La base de sondage est constituée des logements recensés en 2011, 2010 et 2009 (ces deux dernières années étant très minoritaires (3% des adresses). Dans cette base, sont renseignés, pour chaque logement, son adresse, le nom et le prénom de la personne de référence du ménage.

A partir de cette base de sondage, un échantillon de 4500 logements a été tiré par l'INSEE par sondage stratifié sur les critères suivants:

- les zones d'action enquêteur (ZAE).
- les typologies de communes (rural/urbain, taille des communes).

Au sein de chaque ménage sélectionné par l'INSEE, une seule personne devait être sélectionnée aléatoirement. Le sondage est donc stratifié à deux degrés (ménages puis individus). Le poids de sondage est obtenu à partir de la probabilité de sélection du logement (fournie par l'INSEE) et de la probabilité de sélection d'un individu au sein de son logement (fournie au cours du protocole).

L'échantillon initial fourni par l'INSEE a été subdivisé en une partie principale (3500 logements) et deux réserves, à utiliser si l'exploitation de la partie principale ne permettait pas d'atteindre l'objectif initial de 1500 panélistes.

- Le recrutement du pilote

Afin de tester différents modes de contact, le recrutement du pilote a été réalisé selon une procédure séquentielle, du mode le moins couteux au plus onéreux : courrier postal, puis téléphone et enfin face-à-face. Lors de la

première phase de recrutement (réalisée par l'Ined entre juin et septembre 2012), la partie principale de l'échantillon INSEE, soit 3500 logements, a été exploitée. Le protocole a reposé sur l'envoi de courriers d'invitation à participer au panel, à destination de la personne de référence des ménages. Deux types de relance ont ensuite été réalisés: relance téléphonique dans le cas où un numéro était connu (45% des cas) et relance postale sinon.

Il a ensuite été fait appel à l'institut de sondage TNS-Sofres pour la seconde phase du recrutement. Celle-ci a porté sur une large partie des ménages non contactés ou qui avaient refusé de répondre lors de la première phase et sur les échantillons de réserve. Elle a eu lieu pour partie par téléphone, pour partie en face-à-face. Au total, 3629 ménages étaient concernés dont 2642 adresses de l'échantillon principal

Au sein de chaque logement, un individu parmi les éligibles était sélectionné par tirage aléatoire simple. Si la personne sélectionnée acceptait de participer, elle devait impérativement signée une convention précisant le cadre du prêt et les conditions d'utilisations de la tablette au sein du panel. A la réception et l'activation de la tablette par le panéliste, une formation téléphonique à l'utilisation de la tablette lui était proposée et fournie gratuitement via l'opérateur de téléphonie mobile.

Résultats globaux

- Recrutement dans le panel

Au terme de l'exploitation complète des 4500 adresses de l'échantillon fourni par l'INSEE, 1039 individus ont accepté de faire partie du panel et ont signé la convention. Cependant, seuls 923 individus ont répondu à l'enquête annuelle dans laquelle il s'agit de collecter les informations socio-démographiques des panélistes. En effet, une partie est entrée dans le panel après l'enquête annuelle, et une autre partie n'a jamais utilisé sa tablette pour répondre aux enquêtes.

Le recours à un sondage à deux degrés (ménages puis individus) implique de fournir des indicateurs de participation au niveau ménage et individu. Suivant les recommandations fournies par l'American Association for Public Opinion Research (AAPOR : www.aapor.org) et notamment celles adaptées aux panels internet par Callegaro [4], il faut distinguer plusieurs issues à chaque niveau : contact, refus, acceptation (dite aussi participation).

Notons que l'éligibilité est parfois délicate voire impossible à déterminer (dans les cas où aucun contact avec des membres du ménage n'a eu lieu) : pour tous ces cas, on applique le taux d'éligibilité observé sur les unités pour lesquelles l'éligibilité est connue.

Avec ces conventions, le taux de panélistes inclus initialement (n=1026 parmi les 1039) est de 26%, le taux de refus de 43% et le taux de non contact de 31%. Ces taux ne sont pas plus élevés pour la partie du panel n'ayant fait l'objet que d'un recrutement en face-à-face par l'institut de sondage.

- Evolution de la participation aux enquêtes

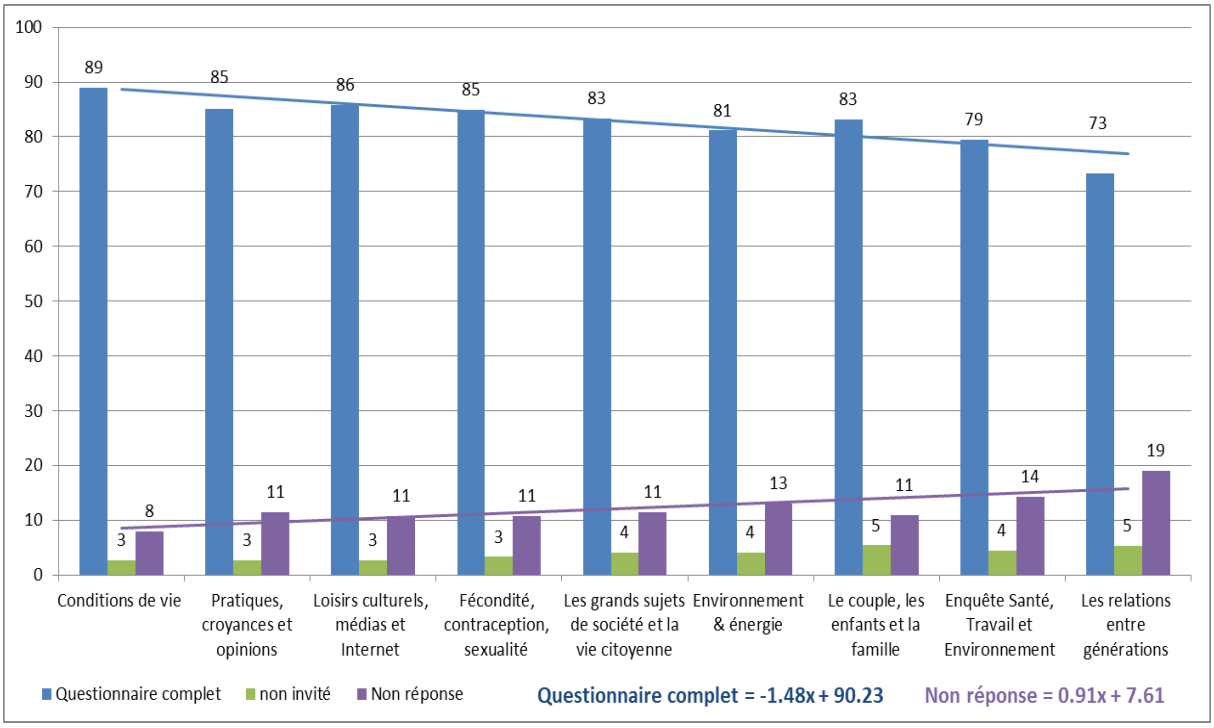
A l'heure actuelle, 12 enquêtes ont été administrées et 3 sont encore en cours (leur durée de collecte ayant été prolongée). Nous présentons donc ici les résultats pour les neuf premières[1]. L'évolution de la participation au cours de ces neuf enquêtes achevée est la suivante (figure 1). Elle montre clairement une baisse continue de la participation au fil du temps (1,48 points en moins à chaque enquête, en moyenne), corolaire d'une augmentation des non-réponses (0,91 point en plus à chaque enquête, en moyenne). La proportion de non invités augmente aussi, mais faiblement, pour rester aux alentours de 5%.

Les panélistes sont en principe invités à répondre à chaque enquête mais ils peuvent avoir un problème de tablette (dysfonctionnement, désactivation de la part des gestionnaires de panel) ou bien être indisponibles (vacances, maladie, etc.). Ils peuvent également ne pas répondre malgré les relances. La désactivation définitive et donc la sortie du panel intervient après des non réponses immotivées trop fréquentes (généralement trois successives).

L'attrition est donc relativement importante, puisqu'elle est d'environ 16 points en 9 enquêtes successives administrées sur une période de 10 mois.

[1] 1 : Conditions de vie ; 2 : Pratiques, croyances et opinions ; 3 : Loisirs culturels, médias et Internet ; 4 : Fécondité, contraception, sexualité ; 5 : Les grands sujets de société et la vie citoyenne ; 6 : Environnement & énergie ; 7 : Le couple, les enfants et la famille ; 8 : Enquête Santé, Travail et Environnement ; 9 : Les relations entre générations ; 10 : Vos conditions de vie en 2014 ; 11 : Les élections municipales ; 12 : Questions générales sur la société française

Figure 1 : Evolution de la participation dans ELIPSS.



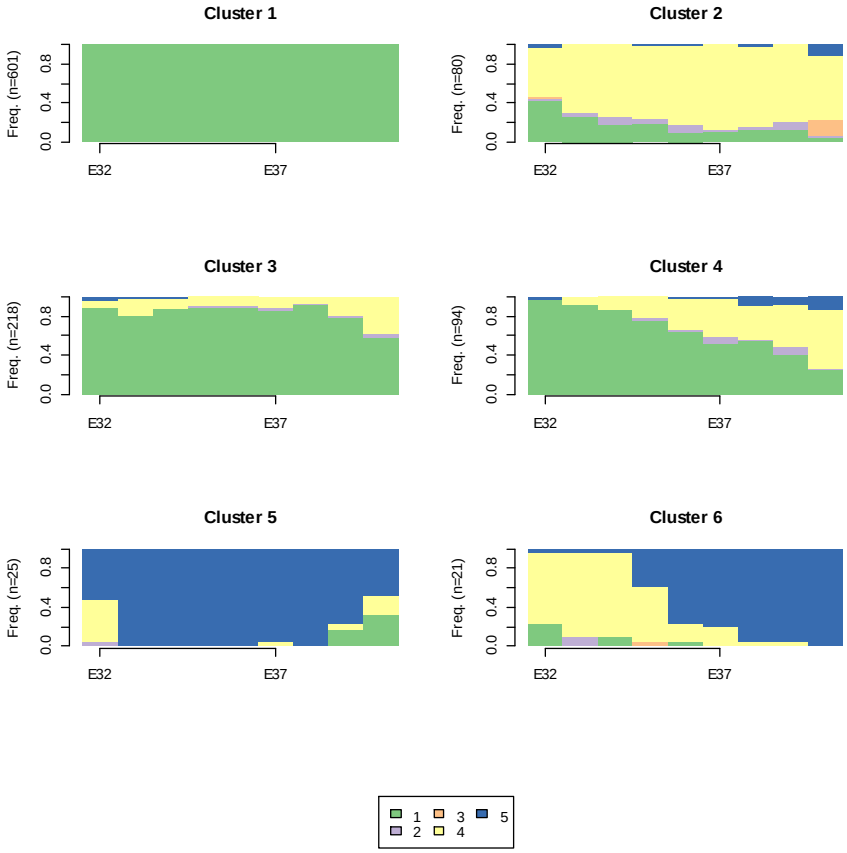
Note : les statuts enregistrés à chaque enquête sont : questionnaire complet, non invité, non réponse, désactivation. Pour plus de lisibilité, ce dernier a été omis de la représentation.

- Différents types de participants

Toutefois, cette évolution globale masque des comportements très différents au niveau individuel. Ainsi, en utilisant les données individuelles de participation à chacune de ces neuf enquêtes, et en procédant à une analyse de séquences (Optimal Matching : [5]) suivie d'une classification ascendante hiérarchique, nous pouvons observer 6 profils de panélistes distincts (figure 2). Ce travail est encore préliminaire (des analyses de robustesse doivent être menées). Le principal groupe est constitué de 600 panélistes (58%) qui ont participé à toutes les enquêtes. Le second est constitué de 80 individus (8%), qui ont peu participé dès le début (ils se distinguent surtout par des non-réponses), mais dont la participation baisse et qui commencent à être désactivé et exclus du panel. Le troisième groupe de 218 individus (21%) est un groupe de participants assez assidus pour toutes les enquêtes, sauf les deux dernières auxquelles ils ont moins répondu. Le quatrième (n=94, 9%) est constitué d'individus dont la participation a décliné très vite dès la première enquête et qui commencent à se voir désactiver et exclure sur les dernières. Enfin, les derniers groupes sont petits (n=25 et 21, 2% chacun) et distinguent d'individus inclus très tardivement dans le panel et dont la participation est faible mais en hausse sur les deux dernières enquêtes ainsi que des individus non participants désactivés très vite. Globalement, les groupes 1 et 3 totalisent 79% d'individus participant systématiquement ou presque à toutes les enquêtes.

Les membres des groupes (*clusters* dans la figure 2) 1&3 sont plus souvent des femmes (51% vs 30% dans les autres), plus souvent actifs occupés (63% vs 38%) ou retraités (16% vs 5%) et plus souvent diplômés du supérieur (42% vs 26%). Ilsrésident plus souvent en habitation individuelle (62% vs 51%, $p<0.001$) et moins souvent en HLM (13% vs 24%, $p<0.001$), et moins souvent en Zone urbaine sensible (ZUS) (4% vs 10%). En revanche, les types de commune (centre, multipolaire, rurale ou non) ainsi que la taille des unités urbaines dans lesquelles elles se trouvent ne sont pas différents de ceux membres des autres groupes. Seuls 3% des individus n'y ont pas répondu à l'enquête sur les conditions de vie, contre 41% dans les autres.

Figure 2 : typologie de la participation dans ELIPSS



- Qualité de l'échantillon de répondants à la première enquête annuelle

Sur le plan des variables individuelles, l'échantillon apparaît assez peu déformé pour ce qui est de l'enquête d'inclusion, traitant des conditions de vie, comme le montre le tableau 1. Les populations suivantes apparaissaient sous-représentées : les 18/24 ans, les 65/75 ans, les personnes nées à l'étranger, les personnes titulaires d'un diplôme inférieur au baccalauréat, ou encore les personnes résidant dans la région parisienne, bassin parisien, le Nord et l'Est. Les populations les plus surreprésentées sont celles des français de naissance, des très diplômés, et les personnes résidant plutôt dans les régions Centre, Sud et Ouest.

Tableau 1 : structure de l'échantillon de répondants ELIPSS

		INSEE RRP[1]	ELIPSS répondants enquête 1	
		18-75 ans	Non Pondéré	Pondéré[2]
Variable	**Modalité**	**%**		
Sexe	Hommes	48.6	48.0	48.6
	Femmes	51.4	52.0 ns	51.4 ns
age	18-24 ans	12.0	8.2	12.0
	25-34 ans	17.7	19.1	17.7
	35-44 ans	20.0	25.7	20.0
	45-54 ans	19.6	21.8	19.6
	55-64 ans	18.1	16.5	18.1
	65-75 ans	12.6	8.8 ***	12.6 ns
Taille ménage	1	16.7	25.9	19.5
	2	34.9	25.8	25.9
	3	20.1	17.9	22.3
	4	18.0	20.6	23.7
	5+	10.4	9.9 ***	8.7 ***
Statut marital	Marié/en couple	51.1	44.9	49.6
	Seul	36.8	40.1	38.2
	Autres	12.1	15.1 ***	12.2 ns
Nationalité	Français de naissance	88.2	90.9	88.2
	Français par acquisition	5.4	5.2	5.4
	Etranger	6.4	3.9	6.4 ns
Statut d'activité professionnelle	Occupe un emploi	59.7	64.9 ***	55.8
	Etudiant(e) ou en stage	4.6	6.0	8.7
	Chômeur	7.8	7.5	8.3
	Retraité(e)	20.1	15.2	19.6
	Autre situation	8.0	6.5 ***	7.6 ***
Statut d'occupation du logement	Propriétaire (ou copropriétaire)	60.5	58.5	59.0
	Locataire ou Sous-locataire	37.2	35.3	32.6
	Occupant à titre gratuit	2.3	4.8	6.5
	NSP ou NVPR	---	1.4 ***	1.8 ***
Dipl	Aucun/CEP/BEPC	28.4	17.6	28.4
	CAP/BEP	24.9	20.9	24.9
	Bac_bac+2	32.3	36.5	32.3
	Bac+3 et plus	14.4	24.9 ***	14.4 ns

[1] : Recensement rénové de la population

[2] : La stratégie de redressement est présentée dans le paragraphe éponyme. *, **, ***, ns : valeur-test du test du Khi² de Pearson respectivement <0.05, <0.01, <0.001 et non significative.

- Redressement

Nous avons pris le parti, dans un premier temps, de ne proposer que des pondérations transversales pour chaque enquête, afin de retenir une méthode simple et reproductible, en nous restreignant à un calage sur marge. La population cible est celle du recensement de la population, dont est issu l'échantillon initial fourni par l'Insee. A l'heure actuelle, seule la première enquête a été redressée. Le choix des variables de calage a été guidé par la taille restreinte de l'échantillon et le contexte du panel ELIPSS. L'âge, le sexe, le niveau de diplôme et la nationalité sont des choix classiques auxquelles la ZEAT a été ajoutée. Les marges ont été calculées à partir du recensement de la population 2010. Au final, le poids obtenu par calage (méthode du raking ratio) a pour moyenne 4,79, écart-type de 2,58, pour un rapport max/min de 13.5 (max=18,9, min=1,40), et coefficient de variation de 53,99.

Discussion et perspectives

Le panel ELIPSS est innovant à plus d'un titre : par sa nature représentative, son caractère d'infrastructure ouverte dédiée à la recherche en sciences humaines et sociales, et l'usage de l'Internet mobile. Sur ce dernier point, il se distingue de tous les panels représentatifs connectés existant à ce jour.

Le pilote a également montré que le recrutement séquentiel en plusieurs modes (courrier, téléphone et face-à-face) était complexe à gérer, nécessitant de nombreux échanges avec la société de sondage et la gestion de nombreux fichiers. Si le panel se poursuite au-delà du pilote, le recrutement de 4000 nouveaux membres pour atteindre la taille finale de 5000 panélistes nécessitera un effort considérable. Conformément à ce qui se dégage de l'analyse des résultats des différentes étapes du recrutement du pilote, celui du futur panel sera confié à un institut de sondage privé pour être effectué en face-à-face lorsque cela est possible, afin de maximiser les contacts et la participation tout en accélérant le processus de recrutement (moins d'échanges de courrier, démonstration de la tablette et signature immédiate de la convention, etc.) Toutefois, la définition du cahier des charges et le suivi du terrain seront des tâches fondamentales, compte tenu des enjeux.

Un soin particulier est actuellement accordé à la définition des stratégies de relance et continueront d'être affinées dans le panel final. Celles-ci pourront s'appuyer sur l'analyse des parcours de participation des enquêtés qui sera complétée et enrichie au fur et à mesure.

Enfin, le rythme mensuel d'enquêtes impose des efforts très importants à l'équipe de production dans la programmation des questionnaires, les tests et les échanges avec les équipes de recherche, le tout dans des calendriers très contraints.

Le panel pilote était pour l'instant réservée aux chercheurs des institutions membres de DIME-SHS. L'ouverture du panel définitif en 2015 aux équipes de recherche externes nécessitera également la définition d'une politique tarifaire spécifique qui reste à imaginer, afin de parvenir à l'autonomie financière, comme le stipule le cahier des charges des équipements d'excellence.

Références

1. Gombault, V., *L'internet de plus en plus prisé, l'internaute de plus en plus mobile.* Insee Première, 2013(1452): p. 1-4.
2. Scherpenzeel, A.C. and M. Das, *"True" Longitudinal and Probability-Based Internet Panels: Evidence From the Netherlands.*, in *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies*, M. Das, P. Ester, and L. Kaczmirek, Editors. 2011, Routledge: London. p. 77-104.
3. Blom, A.G., C. Gathmann, and U. Krieger, *Setting Up an Online Panel Representative of the General Population: The German Internet Panel.* Field Methods, 2013. **27**(4).
4. Callegaro, M. and C. Disogra, *Computing response metrics for online panels.* Public Opinion Quarterly, 2008. **72**(5): p. 1008-1032.
5. Gabadinho, A., et al., *Analyzing and Visualizing State Sequences in R with TraMineR.* Journal of Statistical Software, 2011. **40**(4): p. 1–37.