Données, infrastructures et méthodes d'enquêtes en sciences humaines et sociales

Bilan et perspectives scientifiques - 2014

SOMMAIRE

Introduction

Données, Infrastructures et méthodes d'enquêtes en sciences humaines et sociales (DIME-SHS) est un équipement qui vise à combler le retard accumulé par les sciences humaines et sociales françaises en matière de méthodologie d'enquêtes et de collecte de données[1]. DIME-SHS est un Survey research centre qui tire profit des nouvelles technologies pour offrir des outils à la communauté scientifique des sciences sociales pour collecter et diffuser des données. DIME-SHS s'organise autour de trois instruments :

- DIME Quali : constitution d'une banque d'enquêtes qualitatives afin de dynamiser l'analyse secondaire ;
- DIME Quanti : collecte de données quantitatives par questionnaire assortie de protocoles innovants (panel Internet mobile et centre d'appels téléphoniques) ;
- DIME Web : collecte et analyse des expressions spontanées sur le web.

DIME-SHS a été retenu en 2011 dans le cadre du premier appel à projets « équipement d'excellence » (équipex) du programme des Investissements d'avenir. Coordonné par Sciences Po, DIME-SHS s'appuie sur l'expertise d'un consortium d'institutions de recherche et d'enseignement supérieur : l'Université Paris Descartes, l'Institut national d'études démographiques (Ined), l'École des hautes études en sciences sociales (EHESS), l'école d'ingénieurs Telecom ParisTech, le Groupe des écoles nationales d'économie et de statistique (Genes), le Réseau Quetelet (Très grande infrastructure de recherche Progedo) et EDF recherche et développement (EDF R&D). DIME-SHS a obtenu un financement de 10,4 millions d'euros pour la période 2011-2019.

La gouvernance de DIME-SHS est organisée autour de trois instances. La Coordination exécutive du projet a en charge la bonne marche opérationnelle du projet. Elle est composée du coordinateur du projet, Laurent Lesnard, de la responsable du service informatique du CDSP, Geneviève Michaud, de la secrétaire générale du CDSP ainsi que des responsables des trois instruments : Guillaume Garcia (quali), Anne Cornilleau et Anne-Sophie Cousteaux (quanti) et Mathieu Jacomy (web). Le Comité de pilotage, où siègent les directeurs ou présidents des institutions du consortium, prend des décisions relatives à la direction globale du projet. Enfin, le Conseil scientifique, composé d'experts indépendants internationalement reconnus, formule des orientations scientifiques pour la Coordination et le Comité de pilotage (pour plus de détails, voir p. 53). Au moins trois des membres du Conseil scientifique doivent être spécialistes des questions d'éthiques.

[1] Silberman, Roxane. 1999. *Les sciences sociales et leurs données*, http://www.education.gouv.fr/cid1925/les-sciences-sociales-et-leurs-donnees.html

DIME Quali

Présentation de l'instrument

La création de l'instrument a été motivée par la nécessité de doter la France d'une banque d'enquêtes qualitatives en science politique et sociologie. Cette nécessité a été soulignée par plusieurs rapports à la fin des années 1990 et au début des années 2000. L'ambition originelle est de placer la France au niveau d'autres pays européens, anglo-saxons notamment, en matière de capitalisation des données d'enquête menées selon des méthodologies qualitatives (entretiens, observations, etc.). Les objectifs qui motivent cette création sont, dès le départ, multiples. La mise à disposition des matériaux vise à donner à la communauté scientifique les moyens d'exploiter plus complètement la richesse d'enquêtes souvent sous-analysées, ou permettre de mener plus facilement des comparaisons (dans le temps, dans l'espace, ou entre groupes sociaux). Elle vise également à améliorer les conditions de formation à la recherche via l'enseignement des méthodes, à partir de données réelles et éprouvées, selon le modèle du « data in the classroom ». Enfin le partage des enquêtes revêt une finalité directement épistémologique : en favorisant la transparence sur les pratiques de terrain et la mise en œuvre des méthodes, il s'agit aussi bien de renforcer la scientificité de la démarche qualitative que de favoriser l'histoire des sciences sociales.

Ces objectifs sont difficilement atteignables sans équipement adéquat, et la mise en place de l'instrument entend répondre directement à ce besoin. Les objectifs visés en priorité sont d'encourager l'analyse secondaire (ou réutilisation) des enquêtes archivées, et d'accompagner la formation des futurs chercheurs. En effet ce sont ces objectifs qui ont l'impact le plus direct et le plus immédiat sur les pratiques de recherche, et au-delà le paysage académique. Les objectifs à finalité davantage épistémologique ou historique sont bien sûr importants mais ne pourront être atteints qu'à plus long terme. Pour cette raison, nous avons construit et paramétré l'instrument prioritairement autour des objectifs à court terme, ce qui nous a conduits à faire des choix spécifiques concernant l'instrument. Nous avons volontairement opté pour la création d'un outil scientifique qui met des enquêtes à disposition prioritairement pour qu'on puisse les réutiliser à brève échéance. Dit autrement nous ne nous positionnons pas comme un service d'archives destiné à l'histoire ou l'épistémologie des sciences sociales. L'instrument est donc organisé selon une logique par enquête, c'est-à-dire que ne seront pas archivés des fonds de laboratoires ou des fonds de chercheurs, mais bel et bien des enquêtes.

Sur le plan scientifique, nous avons suivi une ligne directrice visant à construire et à proposer aux chercheurs secondaires des outils permettant de comprendre le contexte de réalisation des enquêtes, ceci afin de leur donner les moyens de les réutiliser valablement. Ce souci s'est traduit par la mise en œuvre de deux grands principes, que nous détaillerons plus loin. D'abord, nous collectons et mettons à disposition les données brutes puis nous nous efforçons de recueillir et d'intégrer le plus possible toute la documentation permettant de renseigner le processus de recherche ; c'est en ce sens que nous diffusons des enquêtes. Ensuite, nous réalisons et offrons à la consultation des chercheurs secondaires une « Enquête sur l'enquête ». L'Enquête sur l'enquête vise à donner les moyens de retracer le processus d'enquête originel, afin de réduire le risque

d'usages décontextualisés des données. En outre, chaque enquête est organisée et présentée comme un mini site web. L'idée d'un « site-enquête" est de faciliter la navigation entre les différents documents archivés et de favoriser ce faisant la compréhension de l'ensemble formé par les matériaux de l'enquête. L'instrument est donc conçu pour jouer le rôle d'une boite à outil destinée à aider les utilisateurs à se familiariser avec l'enquête avant de commencer à travailler avec les données

Enfin, nous avons aussi œuvré à toucher un public large de chercheurs secondaires, d'enseignants et d'étudiants. Nous avions dès le départ l'objectif de faire la connexion entre l'instrument et le portail de diffusion des données du réseau Quetelet ; ceci dans le but non seulement d'éviter de renforcer le clivage quanti / quali, mais aussi de garantir une diffusion large et sécurisée des documents constituant les enquêtes.

Le projet a été conçu plusieurs années avant l'appel à projet des Equipex en 2010. Une réflexion avait été menée antérieurement en France autour de l'analyse secondaire des enquêtes qualitatives, en lien avec un questionnement sur les équipements nécessaires pour la capitalisation de ce type de données. Un premier jalon avait été posé à l'occasion d'un colloque organisé à Grenoble en novembre 2005, dont les actes n'ont jusqu'à présent jamais pu être publiés. Quelques années plus tard des tests d'archivage de données qualitatives ont été effectués au CDSP, avec des chercheurs du Cevipof, un des laboratoires de Sciences Po. À partir de 2009, une étude de faisabilité plus générale a été conduite afin de réaliser un état des lieux de la situation sur l'ensemble des plans concernés (scientifique, éthique et juridique, technique, stratégique). Une première expérimentation a été menée à Sciences Po en partenariat avec le CDSP, à partir de 3 enquêtes choisies pour la diversité des problèmes qu'elles posaient en matière d'archivage et de documentation. On notera que ce sont les trois premières enquêtes mises à disposition via l'instrument (cf. infra). Puis l'obtention en 2010 d'un financement de l'ANR pour le projet réAnalyse a permis de commencer à imaginer - et construire - un prototype de site web correspondant aux objectifs précédemment décrits. Il convient de noter qu'à ce stade il ne s'agissait pas déjà de constituer un équipement pérenne destiné à l'ensemble de la communauté scientifique. Il s'agissait avant tout, dans une approche expérimentale, de tester l'utilité d'un certain nombre d'outils à travers plusieurs projets d'analyse secondaire ; ces derniers soulevaient chacun des questions particulières permettant d'identifier les conditions d'archivage et de mise à disposition nécessaires à une réutilisation convaincante des données. L'instrument quali de DIME-SHS est issu de l'ensemble de ces tests, expérimentations, et de la réflexion qui en a découlé. Il ne reprend cependant qu'une partie des options ou solutions imaginées ou testées précédemment, comme on le verra plus loin.

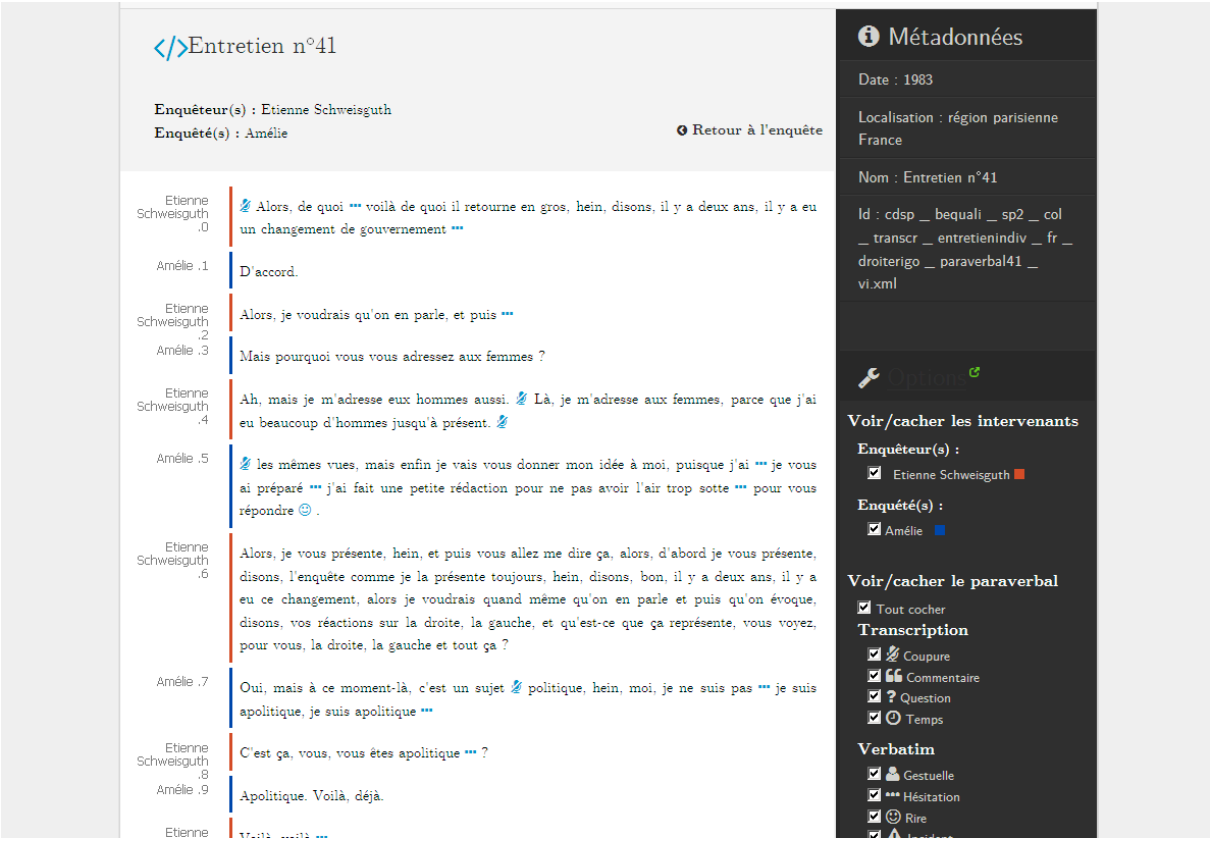
Etat d'avancement du projet

La première phase de construction de l'équipement a débuté (au tout début de l'année 2012) alors que le prototype de site web réalisé dans le cadre du projet ANR réAnalyse était en train d'être achevé. Une première partie du travail a donc consisté à accompagner la fin du développement du site prototype, depuis la livraison de l'application jusqu'au débogage. Une tentative d'utilisation du prototype pour passer à la version de production a été réalisée ; elle a révélé la nécessité d'une adaptation conséquente du site. Or, nous avons dû attendre environ 1 an avant d'être en mesure de pouvoir recruter un développeur dédié. Durant cette période, nous avons dû pallier l'absence de développeur en faisant appel à deux designers qui se sont succédé sur des

missions courtes. A l'issue de cette période nous n'avons pu réaliser qu'une partie des objectifs visés : améliorer la page de présentation des enquêtes, le module de l'enquête sur l'enquête, le moteur de recherche à facettes permettant d'explorer le contenu documentaire de l'enquête, ainsi que l'outil d'édition et d'administration du site. Le développeur recruté à partir de février 2013 a dû ensuite prendre le temps de s'approprier ce qui avait été fait par ses prédécesseurs afin de prolonger l'adaptation du site, dénommé *enQuêtes*. Il a également travaillé sur les visionneuses permettant d'afficher les documents des enquêtes (i.e. des fichiers dans des formats PDF, TEI & CSV). En attendant que l'application de commande Quetelet soit livrée (elle l'a été au printemps 2014) et qu'il soit possible d'y intégrer des enquêtes qualitatives, il a fallu développer une application provisoire servant, pour les utilisateurs, d'interface d'inscription au site et d'interface de commande et de téléchargement des fichiers. Ces contraintes ont fortement contribué à retarder la mise en production du site public. Elles expliquent également que nous ayons dû renoncer à mettre en œuvre plusieurs fonctionnalités qui avaient été testées dans le cadre du prototype : notamment ce que nous avons dénommé les « visualisations », ainsi que la page « Intervenants ». Ces deux interfaces devaient, dans la version prototype, permettre de réaliser en ligne une analyse exploratoire des transcriptions d'entretiens mais aussi des propriétés socio-graphiques des enquêtés. Plus généralement, ces outils de visualisations devaient aider les chercheurs à élaborer des premières hypothèses, conformément aux principes de l'analyse exploratoire des données.

Outre cette première série d'opérations impliquant fortement l'informatique, nous avons dû développer les autres aspects impliqués dans la construction de l'équipement. Une seconde série d'opérations a ainsi été engagée autour de la problématique de l'archivage. L'objectif a d'abord consisté à connecter l'instrument au monde des archives, en établissant des liens avec les acteurs œuvrant au sein des institutions de recherche et d'enseignement supérieur. Concrètement, il s'est notamment agi d'accompagner la création d'un service d'archives à Sciences Po, au plus près du CDSP ; de tisser des liens avec des services d'archives leaders sur la question des archives de la recherche (notamment le service d'archives de l'EHESS) ou des réseaux professionnels (notamment la section Aurore de l'Association des archivistes français). Nous avons œuvré à créer les conditions favorables pour travailler en partenariat avec ce réseau, afin de pouvoir échanger sur nos pratiques, d'améliorer nos procédures, mais aussi d'anticiper la collecte de nouvelles enquêtes. Il s'est agi ensuite de mettre en place des procédures concrètes de conservation des documents collectés. Un plan de classement a été élaboré afin d'organiser les documents constituant une enquête en un ensemble cohérent, comparable d'une enquête à une autre ; on notera que ce plan de classement est utilisé dans le moteur de recherche à facette du site (cf. infra). Des actions ont aussi été engagées pour faciliter l'archivage durable des fichiers numériques, qu'il s'agisse des règles de nommage ou des formats informatiques privilégiés (fichiers texte, audio, tableurs, image, vidéo, etc.). Nous avons également anticipé la question de l'archivage pérenne des enquêtes, en engageant une collaboration avec la Très Grande Infrastructure de Recherche Humanum ; Humanum sert ainsi d'interface entre d'un côté les institutions de la recherche et enseignement supérieur, ou des instances comme l'Equipex, et de l'autre le CINES - l'instance officiellement chargée de la mission d'archivage pérenne des fichiers numériques pour l'enseignement supérieur et la recherche.

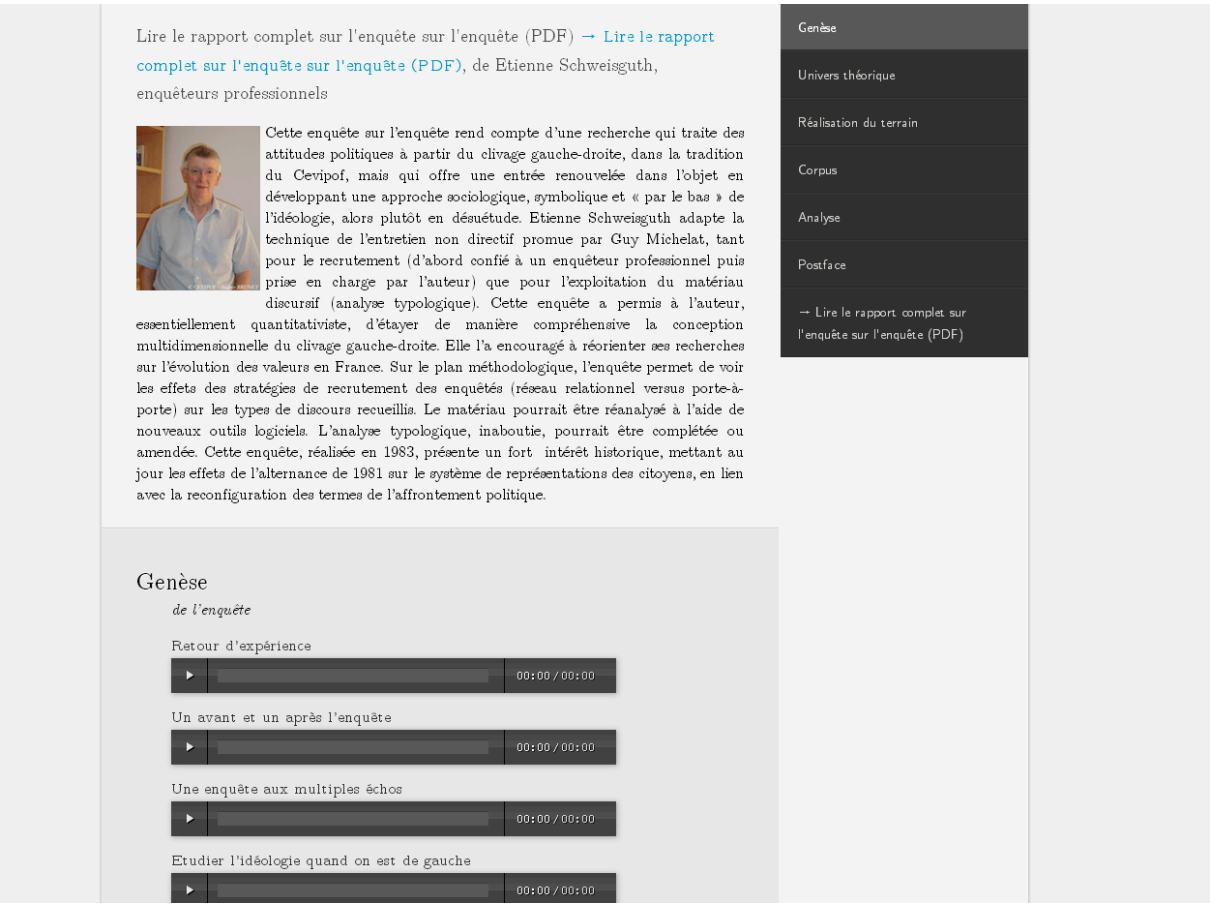
En parallèle, nous avons travaillé à organiser le traitement des documents collectés à des fins de mise à disposition. Nous avons ainsi développé tout un protocole d'encodage des transcriptions d'entretiens en TEI. La structuration des niveaux verbal et paraverbal permet en effet, non seulement, de garantir la pérennité des documents textuels encodés de cette manière, mais aussi d'accroitre les possibilités d'exploration en ligne des documents, et ce de manière interactive, grâce aux outils d'affichage développés sur le site. Pour cela il a fallu développer un dictionnaire de balise TEI adaptées aux transcriptions d'entretiens, ainsi qu'un protocole d'encodage des textes. Il a fallu aussi développer des outils d'affichage (visionneuses) adaptés. Ce travail a été réalisé pour le prototype réAnalyse à partir des enquêtes test ; il a dû ensuite être adapté au site de production (enQuêtes).



Nous avons par ailleurs développé un protocole de numérisation des documents papier, notamment. Nous avons, pour des raisons de ressources et de compétences, choisi d'externaliser cette tâche. Pour ce faire nous avons élaboré un cahier des charges et réalisé un test de numérisation avec un prestataire. Ce travail a pu être mené grâce à la collaboration avec la bibliothèque de Sciences Po, qui a fourni des moyens financiers et humains indispensables au projet.

En parallèle, nous avons engagé un travail de réflexion proprement scientifique, afin d'affiner nos objectifs. Ce faisant il s'est agi non seulement d'améliorer notre efficacité mais aussi de mieux circonscrire ce que nous faisons lorsque nous traitons et mettons à disposition des enquêtes. Nous avons de la sorte voulu nous assurer que nous produisions quelque chose qui soit à la fois utile et valide scientifiquement pour les utilisateurs de l'instrument.

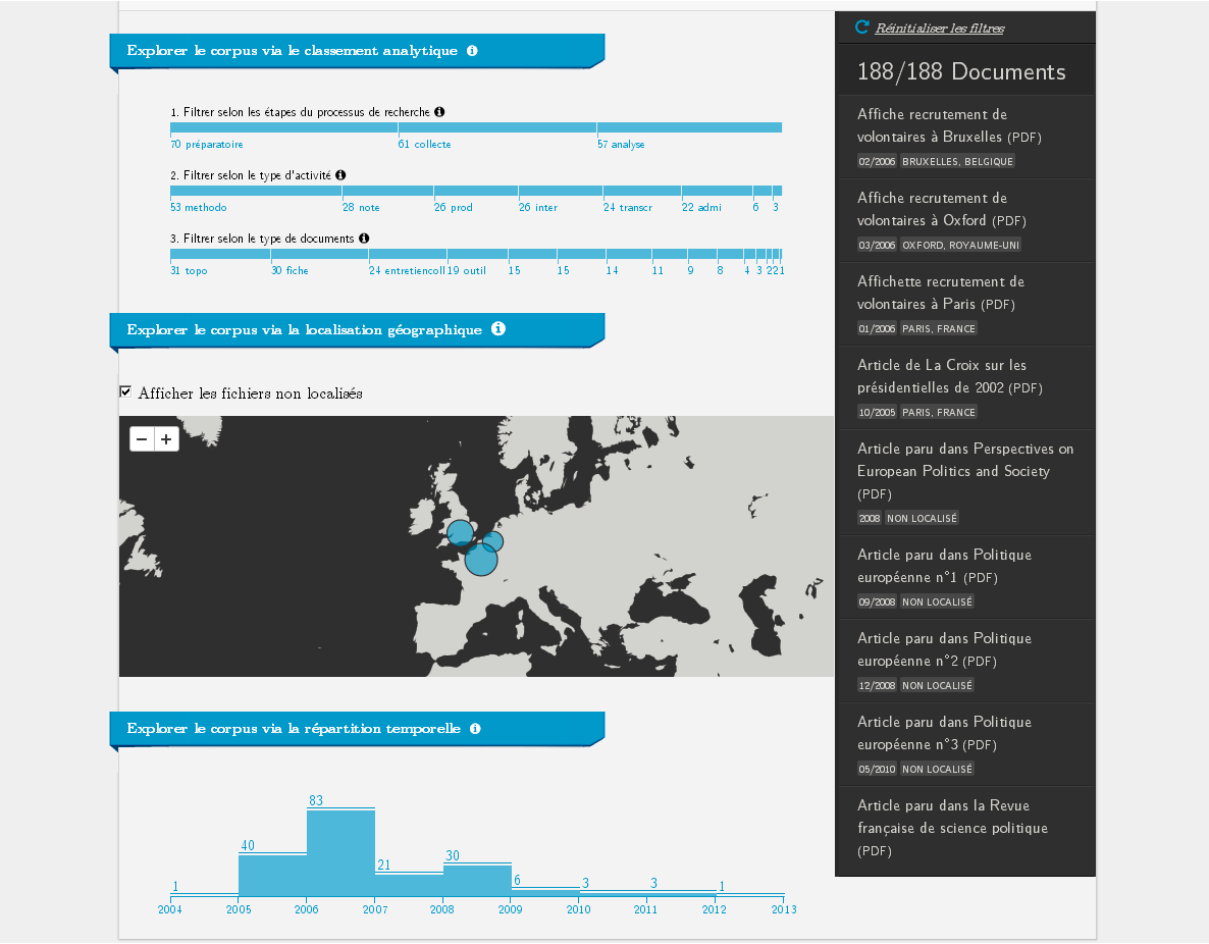
Cet effort a été mené, d'abord, autour de ce que nous appelons l'enquête sur l'enquête. Nous avons sensiblement évolué depuis les premiers tests réalisés pour le prototype ; nous avons notamment abandonné l'ambition de faire, même a minima, une socio-histoire de l'enquête, ni même de faire une analyse de ce que l'enquête a apporté par rapport à l'état de l'art de l'époque. L'enquête sur l'enquête vise essentiellement à restituer, aussi bien que possible, ce que le chercheur sait et pense de son enquête, à fournir le cadre d'intelligibilité qui est le sien par rapport à ce travail. Cela implique de prendre connaissance des documents constituant l'enquête et de réaliser certaines lectures pour s'informer aussi bien sur le contexte scientifique de l'époque que sur le domaine de recherche en question. Mais ce travail est destiné avant tout à préparer le ou les entretiens avec le chercheur premier. Nous ne cherchons donc pas à en savoir plus que lui : nous voulons savoir ce qui est nécessaire pour parler efficacement avec lui, i.e. pour l'accompagner au mieux dans les efforts qu'il fera pour se souvenir de la façon dont les choses se sont passées. Concrètement, l'enquête sur l'enquête est restituée sous deux formes : sous la forme d'un long rapport téléchargeable en PDF (20 à 30 p.) ; sous la forme de chapitres audio qui sont consultables en ligne (ces chapitres résultent du montage de ou des entretiens réalisés préalablement avec le chercheur premier). Le principe de distinction entre ces deux formes est que le texte est construit de manière synthétique, alors qu'on choisit au montage oral les moments les plus vivants, les plus illustratifs du discours du chercheur. Cette forme orale revêt une dimension pédagogique forte, notamment pour les étudiants, c'est-à-dire les publics les moins familiers des enquêtes de terrain. Les deux versions, écrite et audio, sont organisées selon les mêmes six grandes rubriques qui permettent de distinguer et d'articuler les informations nécessaires à la compréhension de l'enquête : sa genèse ; son univers théorique ; la réalisation du terrain ; le corpus constitué et conservé ; l'analyse des matériaux ; et enfin, en guise de postface, la postérité et les potentialités de l'enquête.



Un effort de définition des contours documentaires de l'enquête a ensuite été mené. Il nous a permis de déterminer les types de documents qui nous paraissent nécessaires à la mise à disposition d'une enquête dans les objectifs qui sont les nôtres. Afin de reconstituer le contexte de production de l'enquête, nous privilégions une vision volontairement large des contours documentaires, intégrant les documents préparatoires à la recherche (administratifs, méthodologiques…) jusqu'à l'analyse (brouillons, versions intermédiaires et définitives) en passant bien sûr par les matériaux collectés. Notre démarche de collecte et de mise à disposition des documents ne vise cependant pas l'exhaustivité : il s'agit d'une collecte raisonnée pour laquelle nous cherchons à rassembler uniquement ce qui permettra à d'autres chercheurs d'utiliser les matériaux proposés de façon valide, i.e. en disposant des informations nécessaires à leur interprétation. Cette liste (reproduit en annexe 2 p. 55) sera mise à jour au fur et à mesure de l'archivage et de la mise à disposition de nouvelles enquêtes.

En lien avec ce qui précède, nous avons œuvré à développer des outils de navigation dans le corpus documentaire des enquêtes. Concrètement, chaque document est indexé selon de multiples critères (temporel, géographique, analytique…). L'enquête peut alors être représentée visuellement de différentes manières : la répartition des documents sur une ligne de temps donne à voir la chronologie de l'enquête ; leur répartition sur une carte donne à voir la dimension géographique du processus de recherche ; la navigation à travers les catégories utilisées dans le plan de classement permet de mieux appréhender la structure du corpus documentaire. Ces outils permettent d'opérer des tris et des sélections dans l'ensemble du corpus sur la base de ces critères, en les combinant si nécessaire. Ces outils de navigation sont destinés à faciliter l'appropriation

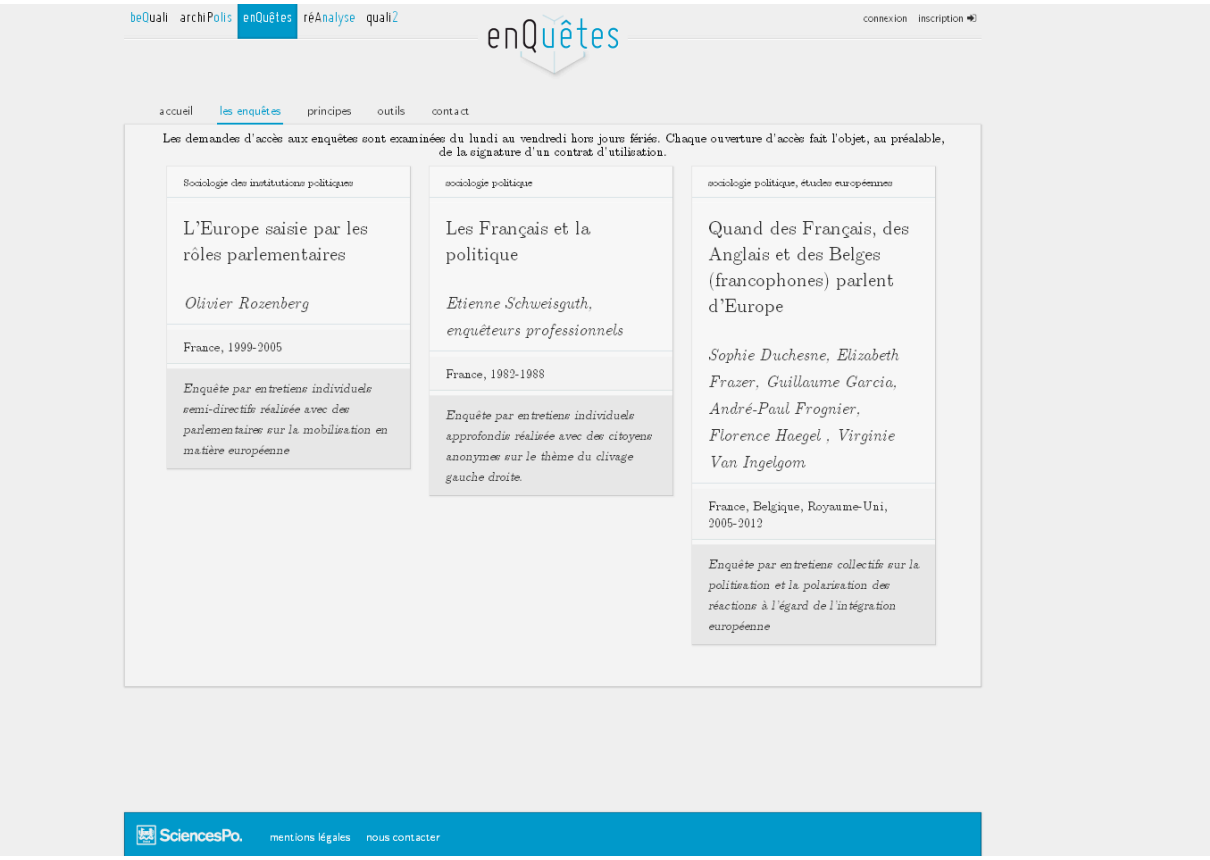
des enquêtes constituées de documents nombreux, diversifiés et dont l'architecture est complexe. Il s'agit d'aller au-delà du simple parcours dans un plan de classement figé, qui enferme le chercheur secondaire dans une arborescence de fichiers. Cette exploration a donc été conçue dans le but de faciliter l'appropriation de l'enquête par l'utilisateur, de lui donner les moyens d'appréhender la composition du corpus documentaire que nous avons rassemblé sans préstructurer le chemin de ses découvertes. Tout cela se passe bien sûr en amont de l'analyse, que le chercheur secondaire mènera comme il l'entendra, avec ses outils habituels, après avoir téléchargé les documents.



L'ensemble de ces activités ont été menées de front depuis le début de l'année 2012. À la date de rédaction de ce rapport, nous avons traité l'ensemble de la chaîne pour 3 enquêtes : la première enquête a été publiée en même temps que la mise en production du site (http://bequali.fr/app), en juillet 2013 ; la dernière l'a été en février 2014. A l'issue de cette phrase, un travail de finalisation de l'instrument – du moins dans sa première version opérationnelle – a été entrepris. Il s'agit de stabiliser aussi bien la partie informatique (correction des derniers bugs apparus avec la publication des trois enquêtes) que nos propres procédures de travail, pour améliorer notre efficacité dans le traitement des enquêtes à venir.

Nous disposons donc d'un catalogue constitué de trois enquêtes :

- **enquête 1** : *Quand des Français, des Anglais et des Belges (francophones) parlent d'Europe* (dir. Sophie Duchesne), 188 documents explorables en ligne, 871 documents accessibles au téléchargement (dont 527 questionnaires) ;
- **enquête 2** : *Les Français et la politique* (Etienne Schweisguth), 73 documents explorables en ligne, 76 documents accessibles au téléchargement ;
- **enquête 3** : *L'Europe saisie par les rôles parlementaires* (Olivier Rozenberg), 142 documents explorables en ligne, 169 accessibles au téléchargement ;



Deux autres enquêtes sont en cours de traitement et devraient être prêtes d'ici à la fin du premier semestre 2014 : il s'agit de l'enquête *La boutique contre la gauche* de Nonna Mayer ; l'enquête *Représentations du champ social, attitudes politiques et changements socio-économiques* de Guy Michelat, Michel Simon et Jean-Marie Donégani. Nous avons également noué des contacts avec une poignée d'autres chercheurs afin de travailleur sur leurs enquêtes au second semestre 2014 (dont par exemple l'enquête *Choisir son école* de Agnès Van Zanten, ou encore l'enquête *La formation des couples* de Michel Bozon et François Héran).

Aujourd'hui, l'utilisation de l'instrument se fait donc sur le catalogue des trois enquêtes. Fin 2013, les statistiques d'utilisation du dispositif étaient les suivantes :

- 34 personnes se sont inscrites en tant qu'utilisateurs
- nous avons eu une dizaine de demandes d'accès aux enquêtes à des fins de recherche (sans compter les membres de l'équipe ni les membres de notre réseau plus large)

- nous avons eu 2 demandes d'accès aux enquêtes à des fins d'enseignement des méthodes. Une première demande dans le cadre d'un cours de méthodologie qualitative à Sciences Po (21 inscrits). Une seconde demande dans le cadre d'une formation à un logiciel d'analyse textuelle réalisée au sein du laboratoire Pacte, à Grenoble (8 inscrits)

La première enquête n'ayant été publiée qu'à l'été 2013 et l'analyse secondaire prenant du temps, aucun projet de réanalyse n'a encore pu être mené à terme. Nous n'avons par ailleurs pas mis en œuvre de véritable politique de communication pro-active autour du projet. Malgré cela, l'instrument commence à être utilisé non seulement par des partenaires du projet réAnalyse mais aussi par des chercheurs extérieurs au projet. Lorsque la publication des deux autres enquêtes en cours de traitement aura eu lieu, nous avons l'intention d'engager une promotion active de la banque, en direction des chercheurs et enseignants, des laboratoires, des instances et réseaux de la recherche.

Nous avons cependant déjà commencé à promouvoir le projet mené jusqu'ici. Nous avons ainsi avancé du côté des publications. Quatre articles rédigés par des membres de l'équipe ont été produits ou acceptés pour publication en 2013 (leur parution étant prévue pour 2014) ; ils mettent en avant une réflexion sur l'équipement et ses implications dans les transformations des pratiques de la recherche en sciences sociales :

- Anne Both, Guillaume Garcia, « Le chercheur, l'archiviste et le webmaster : la polyphonie patrimoniale ? Le cas de beQuali, banque d'enquêtes qualitatives en sciences sociales. », in Bernadette Dufrêne (dir.), *Patrimoines et humanités numériques : quelles formations ?*, Berlin, Lit Verlag, 2014, à paraître.
- Sophie Duchesne, Guillaume Garcia, « beQuali : une archive qualitative au service des sciences sociales », in Marie Cornu, Jérôme Fromageau (dir.), *Les archives de la recherche, pratique des acteurs et enjeux juridiques*, Paris, Editions L'Harmattan, collection droit du patrimoine culturel et naturel, 2014, à paraître.
- Sophie Duchesne, Guillaume Garcia, Anne Both et Sarah Cadorel, « Retour vers le futur : la numérisation des enquêtes qualitatives de sciences sociales entre patrimonialisation et transformation des pratiques scientifiques. », texte mis en ligne le 20/02/2014, consultable à http://humanum.hypotheses.org/147
- Sophie Duchesne, Guillaume Garcia, « Partager les enquêtes en sciences sociales : la révolution numerique *behind closed doors* », *Socio*, dossier "Les sciences humaines et sociales à l'ère du numérique : approches critiques", article soumis le 28/02/2014.

Nous avons également assuré un grand nombre de présentations publiques autour du projet, dans le cadre de séminaires, colloques, conférences, etc. Parmi celles-ci, notamment :

- organisation d'un séminaire sur l'archivage des enquêtes qualitatives en Europe (Paris, novembre 2011)
- participation aux activités du DDI qualitative working group (Göteborg, décembre 2011)
- Présentation au colloque Les archives de le recherche (Paris, janvier 2012)
- présentation à la conférence annuelle de IASSIST (Washington, juin 2012)

- présentation à la conférence *EDDI13 – 5th Annual European DDI User Conference*, (Paris, décembre 2013)

Par ailleurs, au début de l'année 2014 deux postdoctorants (15 et 18 mois) ont été recrutés au CDSP dans le cadre d'un financement Idex Sorbonne Paris Cité. Chacun va réanalyser une enquête, et produire une réflexion autour des conditions méthodologiques de l'analyse secondaire.

Perspectives

Les perspectives s'organisent autour de trois grands axes : accroître et diversifier le catalogue d'enquêtes ; améliorer l'outil de mise à disposition des enquêtes ; approfondir la recherche méthodologique autour de l'instrument.

1-Notre objectif prioritaire à partir de 2014 va être d'accroître et de diversifier le catalogue d'enquêtes mises à disposition.

La réalisation de cet objectif va dépendre de notre capacité à traiter plus d'enquêtes. Elle va dépendre également de l'organisation qui va se mettre en place pour inventorier et collecter les enquêtes. Cette problématique n'avait pas été directement prévue dans le projet initial, tel qu'il a été déposé en 2010 ; les tâches qui en dépendent ne sont donc pas financées dans le cadre de l'Equipex. Il s'agira ici de prospecter de manière active, de constituer un réseau de correspondants dans les services d'archives et dans les laboratoires, afin d'assurer un flux continu d'enquêtes. Une difficulté importante provient du fait que ces opérations vont se dérouler dans un univers académique en sciences sociales marqué d'une part par l'absence de la culture du partage ou de la mise en visibilité des enquêtes qualitatives, et d'autre part par le sous-financement des activités liées à la préservation des archives scientifiques. Néanmoins l'Equipex est déjà structurellement lié au réseau de laboratoires de science politique organisé au sein du consortium archiPolis (lui-même organisé au sein de la TGIR Huma-Num). Il faudra consolider ces liens, sans se limiter à ce réseau. Il faudra notamment développer les liens avec les laboratoires de science politique ou de sociologie inscrits dans le périmètre de l'Equipex. Il faudra également se rapprocher davantage du réseau des archives de l'enseignement supérieur et de la recherche. Ici il convient de souligner que rien ne pourra être fait sans la participation active des chercheurs eux-mêmes. En effet les documents qui constituent les enquêtes sont la plupart du temps en possession des chercheurs. Leur contribution active est également indispensable pour la réalisation de l'enquête sur l'enquête. On ne cherche donc pas à reproduire le système à l'anglaise, caractérisé par une obligation de dépôt lorsque l'enquête est financée sur fonds publics. Au contraire on mise sur la liberté des chercheurs de déposer ou pas leurs enquêtes. Cependant on espère un soutien ou un relais de la part d'instances professionnelles : écoles doctorales, associations comme l'Association française de sociologie ou l'Association française de science politique, ou encore l'ANR, pour sensibiliser la communauté scientifique aux enjeux de l'archivage et de la réutilisation. Dans l'idéal, le dépôt d'une enquête devrait être équivalent à une publication majeure dans le processus d'évaluation de la carrière des chercheurs, de manière à inciter ces derniers à partager leurs données. On espère également que la visibilité donnée au chercheur dont l'enquête serait réutilisée et citée constituera une incitation significative.

L'équipe de DIME-SHS/Quali travaillera également à diffuser auprès des chercheurs et enseignants un guide de bonnes pratiques destiné à faciliter le processus d'archivage des enquêtes du passé mais surtout des enquêtes à venir, comme il peut en exister ailleurs (cf. par exemple le Handbook on *Managing and Sharing Research Data* récemment réalisé par UKDA).

Nous veillerons également à diversifier les enquêtes mises à disposition, non seulement du point de vue des objets et thématiques étudiés, mais aussi du point de vue des méthodes mises en œuvre. Dans un premier temps, nous nous sommes focalisés sur les objets liés au politique, entendu très largement, dans une logique pluridisciplinaire. À terme, on voudrait toutefois constituer des ensembles thématiques eux-mêmes diversifiés, qui correspondent à des domaines d'études dotés d'une certaine consistance : études européennes, sociologie du syndicalisme, sociologie du militantisme et des partis, sociologie de l'administration ou des élites, sociologie des institutions, sociologie des médias, sociologie de la famille, sociologie de l'éducation, etc. À l'intérieur de ces ensembles, nous nous efforcerons de représenter la diversité des courants et des approches, ce qui suppose de représenter au mieux la palette des chercheurs et des équipes de recherche qui produisent ces enquêtes.

DIME-SHS/Quali vise à fonctionner comme un instrument sélectif : le rôle de sélection des enquêtes à archiver en priorité sera joué par le CST. Le CST devra donc non seulement se prononcer sur l'intérêt de mettre à disposition une enquête mais également, tant que nos moyens resteront aussi limités, sur un ordre de priorité. Afin de tenir cet objectif, nous avons voulu que les membres du CST, qui sont des experts en majorité extérieurs au consortium, reflètent la diversité des lignes que nous essayons de faire tenir dans ce projet : diversité des compétences scientifiques, archivistiques et informatiques, diversité disciplinaire et diversité méthodologique et épistémologique. Pour ce faire il s'agira d'établir collégialement une série de critères et de procédures d'évaluation transparents, sur la base de quelques principes, comme notamment : le potentiel de réutilisation des enquêtés aussi bien pour de nouvelles recherche que pour l'enseignement des méthodes ; les risques éthiques impliqués par la mise à disposition des enquêtes ; l'objectif de représentation des divers courants théoriques ou méthodologiques de la sociologie et de la science politique. Lors des deux réunions du CST (juin 2012 et mai 2013) l'examen précis de cette question a semblé prématuré aux experts présents. Le CST a considéré que l'équipe DIME Quali devait continuer à avancer dans la construction de l'instrument en travaillant sur les enquêtes lui permettant de faire le tour des problèmes afférents à la grande diversité des enquêtes qualitatives de sciences sociales, et que ce n'est que plus tard que la sélection devrait et pourrait être organisée via le CST. On notera qu'un CST doit être organisé dans le courant du mois de juin 2014, postérieurement donc à la remise du présent rapport. La question des critères de sélection des enquêtes pourrait être alors à nouveau posée.

2- Un deuxième objectif important va consister à améliorer l'outil de mise à disposition des données.

Dans l'immédiat, il va s'agir d'intégrer le portail Quetelet via l'application de commandes récemment mise en place (mars 2014), afin de bénéficier des fonctionnalités de référencement des enquêtes, de création et de gestion des comptes utilisateurs, et de commande et de téléchargement des fichiers. Nous en sommes actuellement à une phase de test. Nous espérons pouvoir intégrer le portail d'ici l'été 2014.

Dans un second temps, il s'agira d'améliorer les fonctionnalités du site web de mise à disposition des enquêtes. L'adaptation du site prototype en vue d'aboutir au site de production a permis à l'équipe de prendre la mesure du désajustement entre les ambitions projetées dans le prototype et les possibilités de mise en œuvre à plus grande échelle. Il faudra prendre davantage de temps pour développer les fonctionnalités initialement prévues, en prenant en compte les retours des utilisateurs du site, avant de tester de nouveaux outils.

Il s'agira également de tirer parti d'un travail de mise à plat des métadonnées qui a été engagé courant 2013. En effet, les différents travaux relatifs à l'inventaire, à la mise à disposition et à l'archivage pérenne des enquêtes, ont conduit à une utilisation assez hétérogène des informations de description. Si les systèmes liés à ces différentes finalités doivent communiquer, il est indispensable que les informations qui circulent de l'un à l'autre soient structurées. D'autre part nos différents partenariats nous obligent à garantir l'interopérabilité avec leurs infrastructures. C'est le cas par exemple des services d'archives avec lesquels nous devons être en mesure de dialoguer via l'utilisation de l'EAD, avec le réseau Quetelet dans lequel nous devons être référencés en utilisant la norme DDI, avec le CINES pour lequel nous devons être en mesure de générer un SIP (Submission Information Package), etc. Nous avons donc entrepris d'établir un tableau de correspondance entre les différents standards de métadonnées qui nous sont utiles de manière à pouvoir passer nos informations d'un système à l'autre. Avec ce modèle de données, notre système interne doit être en mesure d'importer et d'exporter des informations qui suivent ces différents standards. De plus il nous a permis de corriger et d'uniformiser les métadonnées utilisées dans l'ensemble de notre chaîne de traitement des enquêtes.

3- Enfin nous avons également pour objectif d'approfondir la recherche méthodologique autour de l'instrument.

Un premier chantier consistera, en prenant en compte les retours des utilisateurs, à mieux savoir si le dispositif de l'enquête sur l'enquête est efficace dans sa forme actuelle - c'est-à-dire est-il vraiment utile pour permettre à des chercheurs secondaires, des enseignants ou des étudiants de comprendre une dynamique de recherche qu'ils n'ont pas eux-mêmes vécue ? apporte-t-il une vraie valeur ajoutée par rapport à la documentation déjà disponible sur l'enquête ? Le cas échéant, il faudra également adapter le modèle existant aux problèmes posés par les nouveaux types d'enquêtes que nous serons amenés à traiter.

Il faudra également intégrer les besoins soulevés par les enquêtes nouvellement collectées, notamment les enquêtes recourant à la méthode ethnographique. Le site de mise à disposition a en effet été élaboré à partir d'enquêtes reposant sur des entretiens. Il faut le faire évoluer pour qu'il soit aussi adapté aux enquêtes nourries d'observations,

de prises de notes et de matériel documentaire et iconographique. Seront aussi probablement mis à l'épreuve nos protocoles d'anonymisation, de définition des contours documentaires, de représentation visuelle de l'enquête, etc.

Il faudra également affronter les problèmes spécifiques posés par la mise à disposition d'enquêtes quanti / quali. Pour ce faire, nous bénéficions d'un cas de figure exemplaire avec l'enquête sur la formation des couples de Michel Bozon et François Héran, qui sera prise en charge par un des deux postdoctorants que nous venons de recruter.

Enfin, il faudra approfondir les usages possibles des enquêtes mises à disposition pour l'enseignement des méthodes. Cela pourra revêtir la forme d'un dispositif particulier de consultation pour les étudiants et les enseignants, de kits pédagogiques adaptés, ou encore d'actions spécifiques de formation à l'analyse secondaire, via par exemple l'organisation de séminaires ou d'écoles d'été.

Principes déontologiques et anonymisation pour la diffusion de données

Afin de lever les incertitudes sur l'encadrement juridique de la collecte, du traitement et de la diffusion des enquêtes, nous avons dû faire appel à un cabinet spécialisé à partir de l'automne 2013. Les contrats élaborés, outre le fait qu'ils sécurisent juridiquement les actions entreprises par DIME-Quali, notamment en termes de responsabilité, visent à mettre en œuvre un double principe de protection : protection des chercheurs qui déposent des enquêtes et protection des enquêtés.

En ce qui concerne l'anonymisation, nous gérons deux cas de figure. Premier cas : les enquêtes composées d'entretiens individuels avec des individus qui n'appartiennent pas à un milieu d'interconnaissance et qui n'ont pas de stature publique ; il suffit ici de supprimer les informations directement nominatives pour réduire drastiquement les risques de reconnaissance. Second cas : le cas d'enquêtes mobilisant des enquêtés qui ont une stature publique telle qu'il serait absurde de vouloir procéder à une anonymisation ; la mise à disposition n'est ici possible qu'à condition d'obtenir le consentement des enquêtés, éventuellement ex post. Entre les deux subsiste une zone grise, qui nécessitera l'intervention du Comité scientifique et technique de DIME-SHS/Quali, voire du Conseil scientifique. Est ici en jeu la nécessité d'une réflexion collective et d'une adaptation du niveau d'anonymisation optimal aux caractéristiques de chaque enquête. En effet, sauf à les aseptiser tellement qu'elles en deviennent sociologiquement inutile, il est souvent difficile, dans le cas d'enquêtes localisées en particulier, ou portant sur un milieu bien défini, de faire disparaître tout ce qui permettrait de retrouver la trace des enquêtés.

Pour protéger les enquêtés, nous anonymisons nous-mêmes les données : nous ne nous contentons pas d'exiger des chercheurs qu'ils le fassent (même s'ils l'ont fait au préalable, nous devons le vérifier). Dans tous les cas, nous dressons un tableau d'équivalence (que nous transmet le chercheur déposant ou que nous créons nous-mêmes), conservé à part, dans un endroit sécurisé du serveur de DIME-SHS, à l'écart des fichiers chargés sur le site web.

Actuellement nous avons adopté une politique d'anonymisation basique qui consiste à :

- anonymiser systématiquement les éléments qui permettent une identification directe des enquêtés (nom, prénom, n° de téléphone, adresse…)

- anonymiser autant que possible les éléments qui permettent facilement une identification indirecte des enquêtés (par exemple : maire d'un petit village à une date précise, etc.)

Sur cette base, nous recommandons d'adopter les conventions d'anonymisation suivantes :

- remplacer les marqueurs d'identification directe par un hyperonyme (du type [nom de la personne], [n° de téléphone de telle personne], etc.), de manière à ce qu'on sache de quel type était l'information supprimée,
- lorsque l'élément à anonymiser est complexe (cas typique : maire de tel village), garder la mention qui paraît la plus significative sociologiquement parlant [(maire) et anonymiser le nom du village [nom du village]
- lorsque l'identité n'est pas masquée, attribuer un pseudonyme cohérent socialement au niveau social, géographique et générationnel (remplacer Marcel par Robert et non par Jean-Edouard).

Ces conventions ont été établies à partir d'enquêtes non ethnographiques ; elles seront très probablement amendées quand nous serons confrontés à l'anonymisation d'enquêtes ethnographiques, qui posent des problèmes plus complexes.

Nous estimons indispensable de conserver les autres informations ; les données ne devant pas être trop appauvries sous peine de perdre leur potentiel de réutilisation. De plus, les enquêtes ne seront mises à disposition qu'auprès de chercheurs (quel que soit leur statut). Nous entendons par là les chercheurs statutaires (du CNRS ou d'autres institutions de recherche publique), les enseignants-chercheurs, les ingénieurs de recherche ou d'étude, les post-doctorants et doctorants, ainsi que les étudiants de master recherche (sous la supervision de leur directeur de recherche) ou de licence étudiants (sous la supervision de l'enseignant). Quand ces utilisateurs signent le contrat de réutilisation pour avoir accès à l'enquête, ils s'engagent à respecter l'anonymat des enquêtés, mais aussi la réputation de l'auteur de l'enquête. C'est aussi en principe une forme de garantie vis-à-vis des enquêtés – mais aussi des auteurs des enquêtes qui nous sont confiées – puisque l'activité de ceux qui vont les réutiliser est censément subordonnée à une déontologie, à un savoir-faire, voire à des formes de contrôle collectif.

Les conventions de réutilisation engagent les réutilisateurs à se soumettre à certaines contraintes : respecter les témoins dans les analyses produites à partir des matériaux, ne pas chercher à lever l'anonymat pendant l'analyse et le protéger lors de la publication, ne pas rediffuser les matériaux à des tiers, etc. Au-delà de la protection des témoins, c'est la garantie même de la participation future des chercheurs à l'alimentation du catalogue d'enquêtes archivées et mises à disposition qui est en jeu. Cela implique de formaliser aussi le respect dû au chercheur déposant : le contrat de réutilisation engage à respecter un principe de « civilité scientifique » (de manière à éviter par exemple les règlements de compte) mais aussi à citer le chercheur premier (ainsi que sa principe publication de référence) pour toute production scientifique opérée sur la base de la réutilisation de l'enquête. Ce faisant, on espère – c'est ce qui s'est d'ailleurs passé dans les autres pays qui ont déjà mis en place de genre de dispositif – produire une incitation positive à déposer et documenter ses enquêtes – ce type d'exposition pouvant être considéré comme une contrepartie légitime au partage des données.

DIME Quanti

Le panel ELIPSS (Étude Longitudinale par Internet Pour les Sciences Sociales) est un dispositif d'enquêtes par internet destiné à la communauté scientifique. Il vise à combler l'absence de moyens d'enquête par questionnaire propres aux chercheurs français en sciences humaines et sociales. Par son échantillon probabiliste et sa finalité scientifique, le panel internet de l'équipement d'excellence DIME-SHS est en effet le premier du genre en France et par sa dimension web mobile, en Europe.

Il s'agit d'un panel internet représentatif de la population résidant en France métropolitaine. Une tablette tactile et un abonnement internet mobile sont fournis aux panélistes sélectionnés aléatoirement à partir du recensement afin qu'ils participent aux enquêtes mensuelles. Ces dernières sont élaborées par des chercheurs et sélectionnées par un comité scientifique et technique.

La première partie est consacrée à la présentation générale du dispositif, notamment des contextes français et étrangers en matière d'enquêtes par internet en population générale. La deuxième partie présente le recrutement du pilote du panel ELIPSS, en particulier, les premiers résultats disponibles sur le déroulement du terrain et sur la représentativité du panel. Les différents aspects opérationnels du dispositif sont ensuite exposés, allant de la sélection des enquêtes à la diffusion des données en passant par la production des questionnaires et la participation mensuelle des panélistes. Enfin, un premier bilan du pilote soulignant les enjeux du développement du panel en 2015 est proposé en conclusion.

Un panel internet pour la recherche

Les enquêtes par internet en population générale

Si les enquêtes par internet présentent des avantages certains, leur utilisation en population générale revêt plusieurs difficultés.

D'une part, on retrouve les avantages des enquêtes par questionnaires auto-administrés. Les coûts de collecte sont réduits, essentiellement par l'absence d'enquêteurs ; les enquêtés peuvent répondre au moment qui leur convient le mieux ; l'absence d'enquêteurs permet également d'aborder des questions plus personnelles (santé, sexualité…).

D'autre part, le mode de passation par internet présente des avantages spécifiques. Il permet d'envisager de nouveaux modes de questionnement qui intègrent des vidéos, du son ou des applications interactives. Par ailleurs, la période de collecte peut être réduite puisqu'il n'y a pas (ou presque) de limite au nombre de personnes interrogées simultanément. De plus, les réponses sont sauvegardées au fur et à mesure de leur collecte.

Pour autant, l'utilisation d'internet pour enquêter en population générale se heurte à plusieurs difficultés qui mettent en cause la représentativité de l'échantillon et l'extrapolation des résultats :

Les enquêtes par internet sont réalisées à partir d'échantillons de personnes volontaires, c'est-à-dire d'échantillons non probabilistes.

- Les personnes n'ayant pas accès à internet sont de fait exclues. Or en France en 2012, une personne sur cinq n'avait pas d'accès internet à son domicile (Gombault, 2013).

Afin de dépasser ces biais, une possibilité est de construire un panel internet à partir d'un échantillon aléatoire de la population. Ceci suppose de recourir à un mode de recrutement hors-ligne, d'inclure les personnes qui n'ont pas accès à internet et de les équiper d'une connexion le cas échéant (Das, Ester and Kaczmirek, 2011). Dans le cas du panel ELIPSS, ceci est réalisé par le recours à un échantillon aléatoire d'adresses tiré par l'institut national de statistique et par la mise à disposition de tablettes tactiles connectées en 3G à tous les panélistes.

Les expériences étrangères

Deux expériences étrangères ont inspiré le projet de panel ELIPSS et deux initiatives allemandes sont développées en même temps que le pilote d'ELIPSS. Des discussions sont également en cours dans plusieurs pays pour mettre en place des projets similaires, notamment en Norvège, au Royaume-Uni[2], ou encore en Europe du Sud (un consortium regroupant Chypre, l'Espagne, l'Italie, la Grèce, le Portugal et la Turquie pourrait être construit dans les années à venir).

Le Longitudinal Internet Studies for the Social Sciences (LISS Panel) de l'institut de recherche néerlandais CentERdata (Université de Tilburg)

Ce panel internet représentatif des ménages néerlandais a été constitué en 2007 selon un plan de sondage probabiliste réalisé en collaboration avec le Centraal Bureau voor de Statistiek (CBS, Institut national de statistique des Pays-Bas). Il repose sur la mise à disposition d'un ordinateur simplifié et d'une connexion à internet aux ménages qui en sont dépourvus. Le Liss Panel est constitué de 5 000 ménages, soit 8 000 individus de 16 ans et plus. Ce dispositif est gratuit et exclusivement dédié à des opérations de recherche.

Le KnowledgePanel de Knowledge Networks aux Etats-Unis

Créé en 1999 par deux universitaires américains, ce dispositif repose sur un échantillonnage aléatoire et sur l'installation d'une connexion à internet chez les répondants qui n'en disposent pas à leur domicile. L'échantillon, tiré aléatoirement à partir d'une base d'adresses de logements, est composé de 50 000 personnes de 18 ans et plus. Contrairement au Liss Panel, ce dispositif est également ouvert aux études commerciales.

Le German Internet Panel (GIP) de l'Université de Mannheim

Ce panel représentatif de la population âgée de 16 à 75 ans résidant en Allemagne repose sur un échantillonnage aléatoire de 1 500 personnes. Les membres du panel sont interrogés tous les 2 mois. À l'instar du LISS Panel, un ordinateur simplifié et une connexion internet sont mis à disposition des personnes qui en sont dépourvues.

Ce dispositif est réservé aux chercheurs de l'Université de Mannheim et les questionnaires sont centrés sur les réformes politiques.

---

[2] http://www.natcenweb.co.uk/genpopweb/

Le GESIS Panel du GESIS Leibniz Institute for the Social Sciences de Mannheim

Ce dispositif se caractérise par le recours à deux modes d'interrogation auto-administrés :

- les personnes disposant d'un accès à internet depuis leur domicile peuvent répondre sur le web,
- les personnes ne disposant pas d'un accès à internet ou ne souhaitant pas répondre sur le web peuvent répondre sur papier aux questionnaires envoyés par voie postale.

L'échantillon tiré aléatoirement à partir des registres municipaux est constitué de 4 000 individus âgés de 18 à 70 ans et résidant en Allemagne. Les enquêtes bimestrielles sont proposées par des équipes de recherche en sciences sociales dans le cadre d'appels à projets et ne peuvent aucunement servir un intérêt commercial.

Le dispositif ELIPSS

Le contexte français

En France, il n'existe pas de service de production d'enquêtes par questionnaire à vocation nationale dédié à la recherche en sciences humaines et sociales.

Pour pallier cette absence, les chercheurs français ont deux solutions. Ils peuvent produire leurs enquêtes par le recours à un institut de sondage, mais cette solution est onéreuse, d'autant plus lorsqu'il s'agit de réaliser des enquêtes à partir d'un échantillonnage aléatoire. Ils peuvent également réutiliser des enquêtes statistiquement représentatives, par exemple les enquêtes de la statistique publique, mais celles-ci n'abordent pas l'ensemble des thèmes et des questions susceptibles d'intéresser les chercheurs.

Par ailleurs, on constate une diminution des taux de participation aux enquêtes, liée essentiellement aux refus de répondre et aux difficultés croissantes pour joindre les personnes à interroger. Or ceci fragilise la qualité statistique des enquêtes.

Ainsi, la mise en place du panel ELIPSS répond à deux objectifs principaux :

- permettre aux chercheurs de mener des enquêtes sur des thèmes qui ne sont pas traités par la statistique publique française,
- affranchir la recherche publique des intermédiaires du privé pour la réalisation d'enquêtes par questionnaire à partir d'un échantillonnage aléatoire tout en diminuant les coûts et le temps de collecte (par la passation auto-administrée sur internet).

La collecte par internet mobile

Le panel ELIPSS se distingue des dispositifs similaires à l'étranger par le choix de l'internet mobile comme mode de collecte principal. Une tablette tactile et un abonnement illimité 3G sont fournis à l'ensemble des panélistes en échange de leur participation. Ainsi, ils peuvent répondre aux questionnaires même s'il ne dispose pas de connexion internet.

Le choix de cette nouvelle technologie s'explique par des raisons à la fois méthodologiques et pratiques :

- il s'agit de tirer parti des possibilités offertes par internet (images, vidéos…) et par la mobilité pour renouveler certaines techniques d'enquête (étude des déplacements, carnets d'enquêtes budget-temps, etc.) ;
- les tablettes offrent de nombreux avantages par rapport aux enquêtes sur ordinateur. Parce que leurs interfaces sont plus intuitives, elles offrent un accès internet simplifié aux personnes peu familières des nouvelles technologies. L'accès web mobile donne également aux panélistes plus de souplesse pour répondre aux enquêtes (choix du moment et du lieu),
- l'outil de collecte est le même pour tous les panélistes, via une application spécifique installée sur la tablette.

Outre les avantages de la tablette comme mode de collecte, ce choix a également été guidé par l'effet incitatif attendu par la mise à disposition d'un tel équipement. Le taux de pénétration était faible en France au moment du recrutement du pilote (9% des ménages étaient équipés en 2012[3]). Il faut toutefois noter la forte progression en seulement deux ans puisque près d'un ménage sur trois est équipé à la fin de l'année 2013.

Le calendrier et l'équipe

L'étude pilote a débuté en 2012 afin de définir la procédure de recrutement, d'affiner la méthodologie, de mettre au point les procédures de gestion de panel et de production d'enquêtes et de développer les outils informatiques (cf. calendrier en annexe 3 p. 57). Pendant cette phase de test, l'utilisation du service de production d'enquêtes d'ELIPSS est réservée aux équipes de recherche du consortium DIME-SHS répondant à des appels à projets.

À partir de 2015, le panel ELIPSS devrait être constitué de 5 000 individus et les appels à projets d'enquête seront ouverts à l'ensemble de la communauté scientifique.

La construction de ce dispositif repose sur une équipe aux compétences variées qui s'est étoffée au fur et mesure du pilote (cf. tableau 1). Cette équipe est localisée dans deux institutions. Le Centre de données socio-politiques de Sciences Po assure la coordination du projet, l'ensemble de la production et la diffusion des enquêtes ainsi que les développements et l'infrastructure informatiques. L'Institut national d'études démographiques (INED) a la charge de la gestion du panel ainsi que du suivi de la qualité statistique du panel.

[3] Enquête GfK / Médiamétrie – Référence des Equipements Multimédia

Tableau 1: l'équipe ELIPSS et son évolution depuis 2012

		2012				2013				2014			
Role	Location	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4
Coordination	Sciences Po	Anne Cornilleau Anne-Sophie Cousteaux									to be recruited		
Panel management (including technical support)	INED		Carmen Calandra Gabrielle Bouchet				Elodie Pétorin (40%)	Marc Sigaud			Patricia Sossa Kevin Boudelle		
Survey management	Sciences Po	Emmanuelle Duwez		Matthieu Olivier						Alexandre Mairot	to be recruited		
Statistics	INED									Nirintsoa Razakamanana			
ICT	Sciences Po	Adrien Ferreira *Daniele Guido *Geneviève Michaud *Jérémy Richard (50%)									to be recruited		

*: staff on the 3 instruments of DIME-SHS

Le recrutement du pilote

La procédure d'échantillonnage

La base de sondage

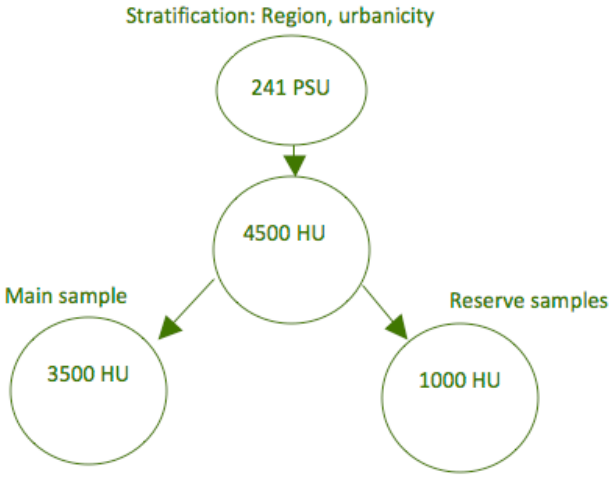
Le panel ELIPSS est un panel aléatoire d'individus résidant en ménages ordinaires au sens de l'Institut national des statistiques et des études économiques (INSEE). Sont donc exclues les personnes sans domicile ou vivant dans des habitations précaires, les personnes résidant dans des collectivités (prisons, maisons de retraite, résidences étudiantes, etc.), et également celles ne maîtrisant pas suffisamment la langue française pour répondre aux questionnaires auto-administrés.

La base de sondage est constituée des logements recensés en 2011[4]. A partir de cette base de sondage, un échantillon de 4500 logements a été tiré par l'INSEE[5] par sondage à

[4] 3% des logements ont été recensés en 2009 ou en 2010.

deux degrés stratifié par région et type de commune (urbain/rural). Le premier degré est un tirage d'unités primaires (PSU) correspondant à des communes ou des groupes de communes, avec un taux de sondage de 1/3[6]. Le second degré est un tirage de logements au sein des 241 PSU. Les 4500 adresses ont été divisées en trois sous-échantillons : un échantillon principal de 3500 adresses et deux échantillons de réserve de 300 et 700 adresses à utiliser au cas où l'objectif des 1500 panélistes n'est pas atteint.

Figure 1: Sampling Design



Au sein de chaque logement sélectionné par l'INSEE, une seule personne est ensuite tirée au sort.

La préparation de la base

La phase préliminaire du travail a consisté à saisir les fiches adresses fournies par l'INSEE et sur lesquelles les noms et adresses des chefs de ménage étaient manuscrites. Une vérification de la cohérence et de la vraisemblance des adresses saisies a été faite manuellement par Internet par les gestionnaires de panel de l'INED. Elle a été complétée par l'institut de sondage pour les adresses des échantillons de réserve. Le format des adresses a également été vérifié automatiquement par la Poste. La base a également été enrichie par la recherche des numéros de téléphone.

Au total, 76 adresses dans l'échantillon principal et 13 dans l'échantillon de réserve ont été considérées comme inexploitables. Par ailleurs, environ 10% des lettres sont revenues avec la mention NPAI (n'habite pas à l'adresse indiquée), ces adresses ont directement reçu la visite d'un enquêteur.

---

[5] Cet échantillon a été fourni gracieusement par l'INSEE à ELIPSS à des fins expérimentales.

[6] L'Échantillon-Maître de l'INSEE est constitué de 567 PSU. La taille effective de l'échantillon du pilote ELIPSS n'était pas compatible avec la mobilisation de l'ensemble des PSU définies par l'INSEE.

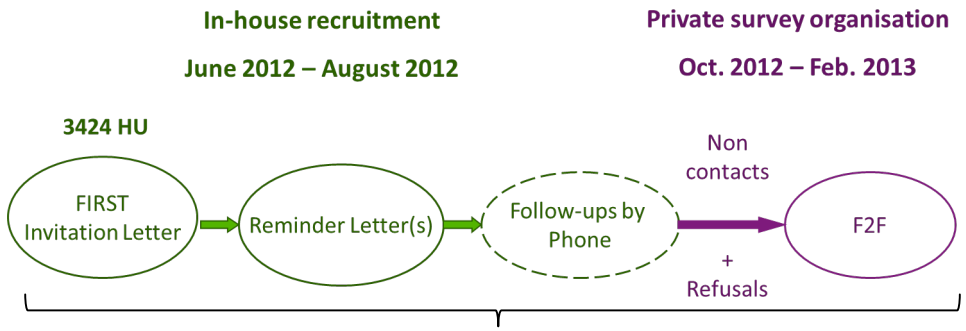
Le déroulement du terrain

Deux procédures différentes d'exploitation des adresses de l'échantillon principal et des échantillons de réserve ont été mises en œuvre. L'échantillon principal a été exploité de façon à étudier l'effet des différents modes de contact (courrier puis téléphone et enfin face à face) et l'exploitation des échantillons de réserve a été réalisée exclusivement par des enquêteurs (téléphone et face à face).

L'échantillon principal

Pour l'échantillon principal contenant 3424 adresses, la procédure de recrutement s'est faite à l'aide de trois modes de contact successifs. Une première invitation à participer a été envoyée par courrier postal en juin 2012 par l'équipe ELIPSS, suivie d'une lettre de relance 15 jours plus tard. Jusqu'à mi-juillet 2012, des relances ont été réalisées par téléphone pour les ménages dont le numéro a pu être retrouvé (environ 50% des adresses de l'échantillon principal). Le recrutement en face à face s'est déroulé d'octobre 2012 à mars 2013 auprès de ménages pour lesquelles aucun contact n'avait encore eu lieu (non-répondants et NPAI) ainsi qu'auprès de la plupart des ménages qui avaient refusé de participer (ces refus représentent 17,5% des adresses qui ont été réexploitées).

Figure 2: Main Sample Recruitment Design

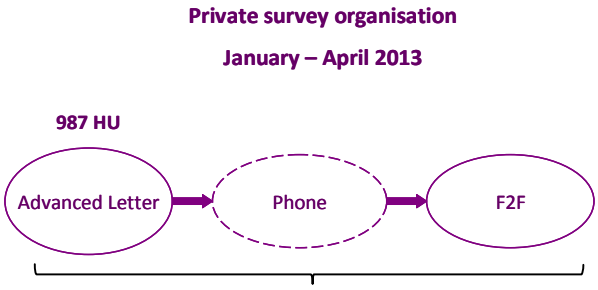


Les échantillons de réserve

A l'issue de l'exploitation de l'échantillon principal, 770 personnes ont été recrutées. Afin de s'approcher de l'objectif visé de 1500 panélistes, il a été décidé en janvier 2013 de confier les 987 adresses exploitables des échantillons de réserve à l'institut de sondage. Lorsque cela était possible, les premières tentatives de contact ont eu lieu par téléphone[7]. Si les 12 tentatives d'appel restaient infructueuses ou si une personne du ménage avait refusé, un enquêteur face à face se présentait à l'adresse. Par ailleurs, les logements sans numéro de téléphone ont fait l'objet d'un recrutement en face à face uniquement. L'exploitation des échantillons de réserve a permis de recruter 256 panélistes supplémentaires.

[7] Un numéro de téléphone a été retrouvé pour 612 logements.

Figure 3: Reserve Samples Recruitment Design



Ainsi, le panel comprenait 1026 individus à la fin du recrutement.

Les actions pour améliorer le taux de recrutement

Plusieurs actions ont été mises en place en vue d'améliorer le taux de recrutement :

- Varier les modes de réponse à l'invitation à participer

Les lettres d'invitation offraient trois modes de réponse aux ménagés sélectionnés. Ainsi, 421 personnes ont renvoyé le coupon-réponse par courrier postal, 248 ont utilisé l'accès qui leur était réservé sur le site internet elipss.fr et 28 ont téléphoné au numéro indiqué sur la lettre. La moitié des personnes qui ont répondu par courrier postal ont accepté de participer alors que c'est le cas de la quasi-totalité de celles qui ont répondu par internet (95%).

- Offrir des bons cadeaux

Le recrutement du pilote a été l'occasion de mener une expérimentation en joignant 2000 bons cadeaux à la première lettre d'invitation envoyée aux logements issus de l'échantillon principal. En s'inspirant des résultats de la littérature internationale et notamment de l'expérimentation réalisée par le LISS panel (Scherpenzeel, 2009), les incitations financières d'une valeur de 10 euros ont été distribuées de manière aléatoire et inconditionnelle lors de l'invitation à participer.

Les ménages ayant reçu un bon cadeau ont plus souvent répondu à l'invitation (pour accepter ou pour refuser de participer) et ont également plus de chances d'avoir accepté de participer (cf. tableau 2).

Tableau 2: Réponse à l'invitation au panel selon l'envoi de bon cadeau

	No answer	**Answered**				**Total**
		Refusal	*Agreement*	*Ineligible*	**All**	
No incentives (n =1424)	67,8	*18,9*	***11,2***	*2,1*	**32,2**	100
Incentives (n = 2000)	62,1	*16,6*	***18,8***	*2,5*	**37,9**	100
Total (n = 3424)	64,5	*17,6*	*15,7*	*2,3*	35,5	100

Note : le test du khi-2 est significatif (p-value <0.001)

L'effet positif des incitations financières sur la réponse et sur l'acceptation n'est certes pas un résultat novateur mais il méritait d'être confirmé en France où cette pratique n'est pas courante dans le domaine des enquêtes scientifiques. En outre, le fait d'avoir reçu un bon cadeau multiplie par 2 les chances d'accepter de participer, toutes choses

égales par ailleurs. La modalisation de l'acceptation à participer au panel montre qu'après l'envoi d'un bon cadeau, le fait d'avoir un téléphone disponible est la deuxième variable la plus déterminante. Dans une moindre mesure, le diplôme du chef de ménage et son âge jouent également.

- Former les enquêteurs

Les 110 enquêteurs ont suivi une demi-journée de formation approfondie, dont le contenu et le déroulement ont été préparés conjointement par l'équipe ELIPSS et l'institut de sondage. Les formations, auxquelles ont pris part des membres de l'équipe ELIPSS, portaient notamment sur la présentation du panel ELIPSS, sur les arguments et les techniques pour convaincre les personnes de participer, sur l'explication des termes de la convention que les personnes doivent signer (cf. section 2.3.1), sur le remplissage du *contact form* et sur la prise en main des tablettes numériques par les enquêteurs. En effet, sur le terrain, les enquêteurs étaient équipés d'une tablette afin de pouvoir en faire une démonstration aux personnes sélectionnées.

- Convertir les refus

Parmi les refus obtenus par courrier ou téléphone, 542 ont été réexploités par l'institut de sondage. En particulier, les refus liés à des questions de confidentialité ou d'anonymat n'ont pas été réexploités. Les raisons évoquées pour les refus à réexploiter sont relativement classiques dans les enquêtes par questionnaire comme le manque d'intérêt pour le panel, le manque de temps pour répondre auxquels se sont ajoutés des refus de personnes déclarant ne pas être à l'aise avec les nouvelles technologie et de personnes ne souhaitant pas avoir un accès à internet ou une tablette. A l'issue de la réexploitation de ces refus, 49 ont donné lieu à une acceptation à participer.

L'entrée des panélistes

La convention

Pour faire partie du panel ELIPSS, les personnes sélectionnées ont signé une convention qui régit les conditions de participation aux enquêtes et d'utilisation de la tablette.

En signant cette convention, les panélistes prennent trois engagements. Ils s'engagent à répondre personnellement et régulièrement aux enquêtes. Ils s'engagent aussi à prendre soin de la tablette et informer l'équipe ELIPSS en cas de casse, de perte ou de vol. Ils s'engagent enfin à restituer la tablette à la fin de leur participation au panel. En échange, ils peuvent librement utiliser la tablette et internet dans le respect de la législation en vigueur. A tout moment, ils peuvent suspendre ou mettre fin à leur participation.

En cas de non-réponse prolongée aux questionnaires ou d'utilisation frauduleuse de la tablette, les panélistes peuvent être exclus du panel.

Outre la mise à disposition gratuite de la tablette et de l'abonnement internet 3G, la convention interdit l'utilisation commerciale des enquêtes et garantit l'anonymat des réponses aux questionnaires.

La prise en main de la tablette

Une fois la convention signée et renvoyée, les panélistes reçoivent une tablette à leur domicile. En page d'accueil, est notamment installée l'application qui permet de remplir les questionnaires.

Les panélistes se sont vus proposer une formation téléphonique pour prendre en main la tablette, découvrir l'application ELIPSS et, le cas échéant, paramétrer le wifi. Ces formations ont été assurées par un prestataire extérieur à partir d'un script défini par l'équipe ELIPSS. Deux tiers des panélistes ont été formés, 30% n'ayant pu être joints par téléphone[8] et 5% ayant refusé la formation. A l'issue de la formation, les formateurs devaient évaluer le niveau d'aisance avec les nouvelles technologies. Ils ont estimé que 20 % des panélistes formés n'étaient pas à l'aise avec la tablette.

La première enquête

Lancée en décembre 2012, la première enquête comportait deux parties :

- Le « didacticiel » devait permettre aux panélistes de se familiariser avec différents types de question et le design des enquêtes ELIPSS.
- Le module « Enquêtes et internet » visait à mesurer l'accès à internet, les pratiques numériques et la participation aux enquêtes avant l'entrée dans le panel ELIPSS.

Administrée jusqu'à la fin du recrutement en avril 2013, 90% des panélistes[9] ont répondu à cette première enquête.

D'après cette enquête, 91% des panélistes disposent d'un accès internet à domicile[10]. Ils sont 79% à utiliser internet tous les jours et 6% à se connecter moins d'une fois par semaine. Cette enquête nous apprend également que 13% des panélistes affirment avoir toujours refusé de participer à une enquête avant le panel ELIPSS. Elle confirme enfin que la mise à disposition d'une tablette tactile a été la principale motivation pour accepter de participer au panel ELIPSS (citée par 62% des panélistes), suivie de la confiance dans les institutions impliquées dans le projet (pour 46% des panélistes). Viennent ensuite l'originalité du projet (37%), l'intérêt pour la recherche (32%) et la mise à disposition d'un abonnement internet (13%).

La représentativité du panel

Il n'importe pas seulement de savoir combien d'individus ont finalement accepté de faire partie du panel ELIPSS et participent régulièrement aux enquêtes, il importe aussi de savoir qui sont ces individus, quelles sont leurs caractéristiques. En d'autres termes, dans quelle mesure le panel est-il représentatif de la population générale ? Dans cette partie, nous donnerons des éléments de comparaison entre le profil du panel issu de l'enquête annuelle et les statistiques nationales.

---

[8] Dans la majorité des cas, ces personnes n'étaient pas relancées pour suivre la formation dans le cas où elles avaient répondu à la première enquête.

[9] Seuls 943 panélistes ont été invités à participer à cette enquête car les dernières personnes recrutées sont entrées dans le panel au mois d'avril 2013 au moment de l'enquête annuelle 2013. En considérant les seules personnes invitées à répondre à « Enquêtes et internet », le taux de complétion est de 99% (cf. tableau 7).

[10] Ces données ne sont pas pondérées.

L'enquête annuelle

L'enquête annuelle ELIPSS a pour objectif de disposer de nombreuses variables socio-démographiques (module signalétique[11]) ainsi que de variables de croisement et d'indicateurs fréquemment utilisés en sciences humaines et sociales (module barométrique[12]). Le questionnaire a été construit en collaboration avec plusieurs chercheurs spécialistes des thèmes abordés et avec les membres du comité scientifique et technique (cf. section 3.1.1). En outre, les questions retenues sont très largement issues d'enquêtes existantes, nationales et internationales. Une majorité des variables issues du module signalétique est systématiquement appariée aux fichiers de données diffusés.

La première enquête annuelle ELIPSS a été administrée en 2013 à l'issue du recrutement du panel, en avril pour le module signalétique et en mai pour le module barométrique. En mars 2014, les deux modules ont été administrés ensemble : le module signalétique a été répliqué dans sa totalité (les questions ont été adaptées de façon à mesurer l'évolution depuis la première interrogation) et une sélection de questions du module barométrique a été reposée, à l'exception de la partie sur les comportements et opinions politiques[13].

Comparaison du profil du panel avec les statistiques nationales

Si la mise à disposition de tablettes tactiles aux membres du panel ELIPSS a pour objectif de faire face au problème de couverture lié aux enquêtes par internet, on constate néanmoins des différences entre les caractéristiques des panélistes et celles de la population cible. Le tableau 3 compare[14] la distribution de plusieurs variables socio-démographiques dans le panel (à partir de l'enquête annuelle 2013) et dans la population des 18-75 résidant en France métropolitaine (à partir des données du recensement de 2010). On observe des distorsions similaires à celles constatées à l'issue du recrutement du LISS panel (Leenheer and Scherpenzeel, 2013 ; Knoef and de Vos, 2009) en termes d'âge et de niveau de diplôme. Les tranches d'âges les plus élevés sont sous-représentées comme attendu, ainsi que les 18-25 ans, groupe toujours difficile à atteindre dans les enquêtes en France. En revanche, les ménages d'une seule personne sont surreprésentés, ainsi que les personnes très diplômées et les personnes occupant un emploi. La répartition hommes-femmes est en revanche très proche de celle de la population. Des analyses multivariées sont en cours pour compléter ces premiers résultats.

---

[11] Ce module traite des questions suivantes : état civil, travail et formation, description socio-démographique du ménage, logement et quartier, revenus et patrimoine.

[12] Ce module s'intéresse aux liens sociaux, aux loisirs et pratiques culturelles, aux croyances et pratiques religieuses, aux comportements et opinions politiques, à l'état de santé, aux comportements de santé et aux habitudes de vie.

[13] Ces questions sont posées régulièrement au panel dans le cadre du projet longitudinal Dynamob.

[14] Les écarts entre les données du recensement et du panel ELIPSS ont fait l'objet de tests du Khi-deux, seules les différences statistiques significatives sont commentées.

Tableau 3: Distribution de quelques variables sociodémographiques de l'échantillon des répondants et comparaison à la population

Variable	Modalité	Recensement INSEE (2010) en %	Enquête annuelle ELIPSS (avril 2013)en %	
		18-75 ans	Non Pondéré	Pondéré
Sexe *	Hommes	48.6	48.0	48.6
	Femmes	51.4	52.0	51.4
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
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
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
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	0	1.4	1.8
Niveau de diplôme *	Aucun/CEP/BEPC	28.4	17.6	28.4
	CAP/BEP	24.9	20.9	24.9
	Bac à bac+2	32.3	36.5	32.3
	Bac+3 et plus	14.4	24.9	14.4

Note : les variables marquées d'une * ont été utilisées pour réaliser le calage sur marge auxquelles il faut ajouter la zone géographique d'habitation.

Une question importante dans la mise en place d'un projet comme Elipss est le fait de fournir un accès à internet à ceux qui n'y avaient pas accès avant leur participation au panel (*offliners*). En comparant les données de l'enquête sur les technologies de l'information et de la communication et le commerce électronique réalisée en 2012 par l'INSEE (tableau 4), on observe que les panélistes vivant dans des ménages déjà équipés d'un accès internet à domicile sont surreprésentés. Ce résultat est comparable dans le cas du LISS Panel et du GIP (Blom, Gathmann, Krieger, 2013 ; Leenheer and Scherpenzeel, 2013).

Tableau 4: L'accès à internet à domicile dans la population et dans ELIPSS

	Enquête TIC INSEE, 2012 (18-75 ans)	ELIPSS
Equipé d'un accès internet à la maison avant ELIPSS	83%	91%
Sans accès internet à la maison avant ELIPSS	17%	9%

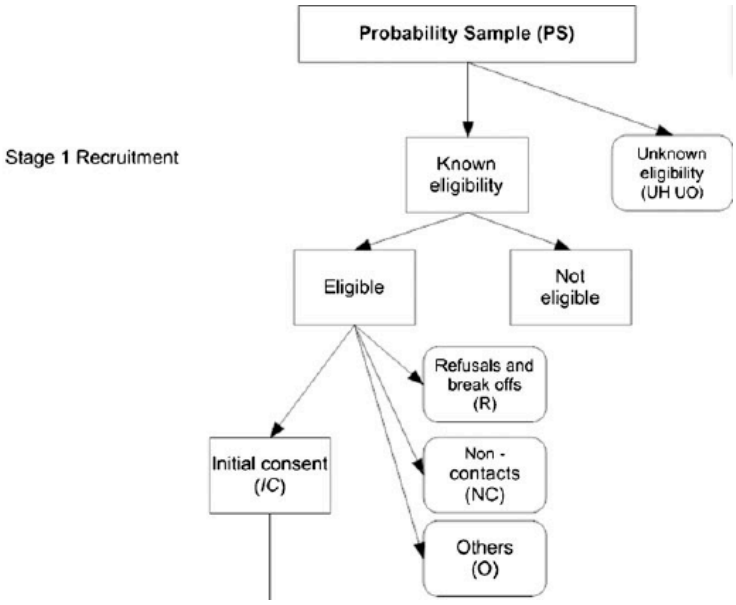
Des analyses sur le profil des *offliners* seront menées en collaboration avec Mélanie Revilla (Université Pompeu Fabra, Barcelona) et Pablo de Pedraza García (Université de Salamanca) à partir de juin 2014.

Résultats globaux

Le calcul du taux de recrutement est un premier élément d'évaluation du pilote. On peut distinguer ici deux types d'accord à participer (*initial consent)* tel que défini par Callegaro et DiSogra (2008). En effet, dans la procédure spécifique aux adresses de l'échantillon principal, une première étape a consisté à demander l'accord du ménage pour décrire la composition des membres du logement[15], puis le second accord intervient au niveau individuel au moment de la signature de la convention de participation au panel. A la différence de la procédure décrite par Callegaro et DiSogra (2008), il n'y a pas eu de *profil/connection stage*, ou plus précisément, elle est confondue avec l'étape dite d'*initial consent* au niveau de l'individu.

[15] Ceci est particulièrement vrai dans le cas de la première phase de recrutement de l'échantillon principal puisque les ménages devaient avoir une démarche volontaire pour renvoyer le coupon-réponse, téléphoner aux gestionnaires de panel ou remplir le formulaire en ligne pour décrire des habitants du logement.

Figure 4: Recruitment Strategy Stage 1 (Callegaro, DiSogra, 2008, p.1012)



Les résultats finaux de la procédure de recrutement du panel ELIPSS sont reproduits dans le tableau 5 ci-dessous. Ils ont été calculés à partir des formules de l'AAPOR, en considérant, comme le proposent Callegaro et DiSogra (2008, p.1018), le RR3 comme taux de recrutement (RECR), tel que définit ci-dessous :

$$\text{Recruitment rate (RECR)} = \frac{IC}{IC + (R + NC + O) + e(UH + UO)}$$

where

IC = initial consent
R = cases directly and actively refusing
NC = noncontacts
O = other cases
UH = unknown if household is occupied
UO = unknown other
e = estimated proportion of cases of unknown eligibility that are eligible.

Tableau 5: Taux de réponse des ménages et taux de recrutement des individus

		ALL households	ALL individuals	*Main sample (individuals)*	*Reserve samples (individuals)*
Eligible	Initial consent (IC)	1352	1026	*770*	*256*
	Refusals (R)	1481	1695	*1318*	*377*
	Non contacts (NC)	0	2	*2*	*0*
	Other cases (O)	65	175	*163*	*12*
Not eligible		570	570	*426*	*144*
Unknown eligibility	Unknown household (UH)[16]	172	172	*89*	*83*
	Unknown other (UO)[17]	860	860	*732*	*128*
	e	0,836	0,836	*0,841*	*0,817*
RR3 = RECR		**0,36**	**0,27**	***0,26***	***0,31***

Ainsi on obtient un taux de recrutement de 27% au niveau de l'individu, avec un taux de réponse de 36% (RR3) au niveau ménage. Par ailleurs, l'exploitation des échantillons de réserve a permis d'atteindre un taux de recrutement légèrement plus élevé que celle de l'échantillon principal. Cette différence tient notamment à un plus grand nombre de ménages pour lesquels l'éligibilité n'a pas pu être déterminée (*unknow eligibility cases)* dans l'échantillon principal, les taux de refus étant similaires (45% dans l'échantillon principal et 46% dans les échantillons de réserve). Ceci s'explique en grande partie par le travail plus approfondi de nettoyage et de vérification des adresses mis en œuvre pour les échantillons de réserve, ce qui a permis de diminuer les non-contacts.

Du projet d'enquête à la diffusion des données

La sélection des enquêtes

Les propositions d'enquête se font lors d'appels à projets. Depuis 2011, un appel par an a été ouvert. Jusqu'en 2013, dans le cadre de l'étude pilote, ces appels à projets étaient réservés aux équipes de recherche membres d'une institution partenaire de l'équipement d'excellence DIME-SHS. Le prochain appel en 2014 sera ouvert à l'ensemble de la communauté scientifique française et internationale.

Le comité scientifique et technique DIME Quanti

Le comité scientifique et technique (CST) DIME Quanti sélectionne les projets d'enquête déposés par les chercheurs lors des appels à projets et examine les demandes

[16] L'adresse n'a pas pu être localisée.

[17] Le ménage n'a pas pu être joint.

d'appariement de données. Il compte 15 membres dont au moins la moitié est affiliée à des institutions extérieures au consortium DIME-SHS. Le CST est composé de chercheurs issus de disciplines variées en sciences sociales (sociologie, sciences politiques, démographie, épidémiologie, etc.) spécialisés en méthodes d'enquête par questionnaire. Il comprend un représentant de l'INSEE, en raison du tirage de l'échantillon par l'institut et de la validation des enquêtes par le Conseil national de l'information statistique (CNIS).

Les critères d'évaluation

La finalité exclusivement scientifique constitue le critère impératif d'éligibilité des enquêtes et l'utilisation commerciale est exclue.

Les projets sont évalués selon les critères suivants :

- la qualité scientifique de la proposition : objectifs, état de l'art, originalité de la proposition, pertinence des échelles ou des indicateurs retenus, résultats attendus ;
- la pertinence de la collecte par le panel ELIPSS : faisabilité (taille de l'échantillon, adéquation des formats des questions à la tablette…), concision du questionnaire (notamment par le recours aux questions de l'enquête annuelle ELIPSS ou d'autres enquêtes ELIPSS déjà réalisées), dimension longitudinale, exploitation des possibilités techniques liées à internet et à la tablette ;
- l'intérêt méthodologique de la proposition : comparabilité des résultats avec d'autres sources, innovations méthodologiques…

Des règles concernant la diffusion des données, le temps d'enquête et les liens avec les panélistes sont décrites dans les appels à projets, notamment :

- Les projets retenus font l'objet d'une convention établissant la copropriété des données entre l'équipe de recherche porteuse du projet et DIME-SHS. Cette convention prévoit le dépôt des données au CDSP et autorise la diffusion des données à la communauté scientifique après une période d'exclusivité d'un an pour l'équipe de recherche porteuse du projet.
- Compte tenu du temps limité d'interrogation mis à la disposition de la communauté scientifique, les propositions d'enquête longue (plus de 30 minutes) ou d'enquête longitudinale doivent être particulièrement justifiées. La durée annuelle cumulée ne peut excéder soixante minutes.
- Aucune incitation financière ne peut être versée aux panélistes. Par ailleurs, les panélistes ne peuvent pas interagir les uns avec les autres.

Les projets soumis

Depuis le premier appel à projet en 2011, 20 propositions d'enquête ont été reçues, parmi lesquelles 11 ont été acceptées par le comité scientifique et technique DIME Quanti.

Tableau 6: Les propositions d'enquête lors des appels à projets

	2011	2012	2013
Nombre de propositions	5	8	7
Nombre d'enquêtes sélectionnées	4	5	2
Nombre d'enquêtes en cours d'évaluation	0	0	2

De décembre 2012 à avril 2014, douze enquêtes ont été administrées au panel ELIPSS. Outre l'enquête sur les pratiques numériques lors de l'entrée dans le panel et l'enquête annuelle ELIPSS, les enquêtes ont abordé des thèmes variés tels que les pratiques culturelles, la contraception, les valeurs et les opinions politiques, l'environnement, le couple, la famille et les relations intergénérationnelles, la santé et les expositions professionnelles (cf. tableau 7).

Deux projets comparatifs ont été acceptés par le comité scientifique et technique en dehors du cadre des appels à projets.

Le premier a été administré au panel ELIPSS en avril 2014. Il s'agit d'un projet porté par Jon Krosnick qui a pour objet de répliquer des expérimentations classiques aux Etats-Unis sur plusieurs panels internet à travers le monde. Les 18 questions visent à étudier la formulation des questions, la tendance à l'acquiescement, les options de non-réponse, l'ordre des questions et l'ordre des modalités de réponse.

Le second sera administré sur ELIPSS en mai 2014 en même temps que sur les autres panels internet probabilistes en Europe : le GESIS panel, le GIP et le LISS panel. Ce questionnaire construit en commun emprunte des questions à de grandes enquêtes comparatives (ESS – European Social Survey, SHARE – Survey of Health, Ageing and Retirement in Europe, PIAAC – Programme for the International Assessment of Adult Competencies, EES – European Election Study). Outre la production de données comparatives au moment des élections européennes, l'objectif est surtout de réussir à organiser cette collecte simultanée sur les quatre panels de manière à pouvoir envisager d'autres projets comparatifs à l'avenir[18].

La production des enquêtes

L'élaboration des questionnaires

A la suite de la sélection du projet d'enquête, des discussions méthodologiques et techniques s'engagent entre l'équipe de recherche et l'équipe ELIPSS. Jusqu'à la mise en production via l'application ELIPSS, le questionnaire évolue en fonction des recommandations du CST, des spécifications techniques, des développements

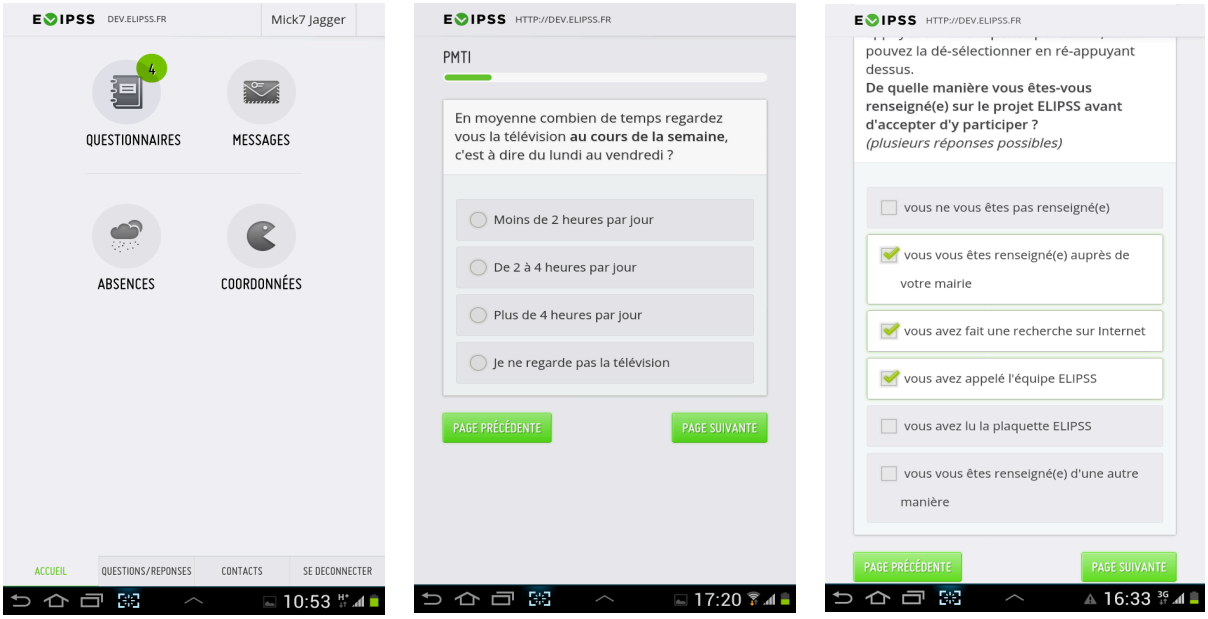
[18] Cette collaboration entre le GESIS panel, le GIP, ELIPSS et le LISS panel s'est également traduite par la soumission en mars 2014 d'un article commun intitulé « A Comparison of Four Probability-based Online and Mixed-Mode panels in Europe » pour un numéro spécial de *Social Science Computer Review.*

informatiques nécessaires, des options choisies pour la non-réponse et des pré-tests réalisés par l'équipe de recherche et l'équipe ELIPSS.

En règle générale, le terrain démarre le 1er jeudi du mois et dure 4 à 5 semaines. Le terrain peut être plus court. Cela a lieu en 2014 pour les enquêtes pré-électorales et post-électorales dont le terrain dépend des dates des élections. Le terrain peut aussi durer plus longtemps. Cela est le cas deux fois par an pour tenir compte des fêtes de Noël et des vacances d'été. Par ailleurs, certains terrains ont été allongés de quelques jours suite à des problèmes techniques.

La tablette comme mode de collecte

Les panélistes répondent aux questionnaires via une application développée en interne et pré-installée en page d'accueil de la tablette (cf. copies d'écran ci-dessous)[19]. La conception des questionnaires et la collecte des données en ligne reposent sur le logiciel Blaise. Développée par Statistics Netherlands, cette solution est destinée à la réalisation d'enquêtes de la statistique publique et est utilisée par la plupart des instituts nationaux de statistiques, y compris l'INSEE. Pour publier l'enquête en ligne avec Blaise IS, il a été nécessaire de développer une feuille de style adaptée à l'écran tactile de la tablette.



Certains projets d'enquête ont été l'occasion de développer des fonctionnalités rendues possibles par la mise à disposition du même équipement mobile, connecté et tactile, à tous les panélistes. Par exemple, les tablettes étant équipées de micro, les panélistes ont pu être invités à enregistrer oralement leurs réponses à des questions ouvertes sur l'environnement. Pour utiliser l'auto-enregistrement dans l'enquête d'octobre 2013, il a fallu préalablement avertir les panélistes dont l'espace de stockage apparaissait insuffisant sur notre logiciel de gestion de flotte de tablettes (*Mobile Device Manager*). La fonctionnalité de glisser-déposer (*drag-and-drop*) a été spécifiquement développée pour

[19] Cette application permet également aux panélistes d'envoyer des messages aux gestionnaires de panel, d'indiquer leurs périodes d'indisponibilité, de modifier leurs coordonnées et d'avoir les réponses aux questions les plus fréquentes.

l'enquête sur les catégorisations et la connaissance du monde social prévue en juillet-août 2014. Ainsi, les panélistes pourront de manière intuitive faire glisser des étiquettes de professions pour créer leurs propres groupes sociaux.

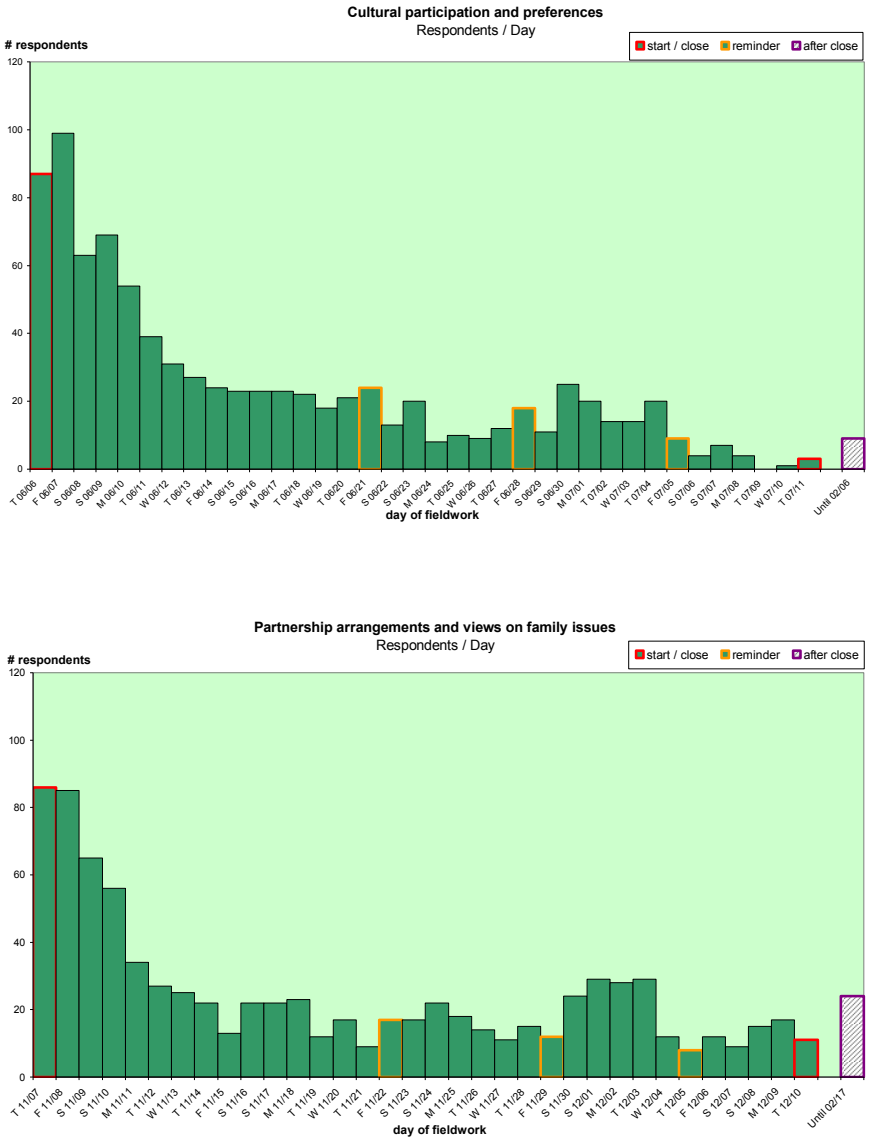
Si l'homogénéité des équipements internet fournis aux panélistes est un atout indéniable d'un point de vue méthodologique (Callegaro, 2010) et d'un point de vue technique, le choix des tablettes implique aussi certaines limites qui tiennent aux problèmes de connexion et à la taille de l'écran. Par exemple, la fonctionnalité d'auto-complétion qui permet de rechercher sa réponse dans une liste (de pays, de communes…) requiert une bonne connexion internet. De même, pour éviter d'avoir à faire défiler le texte verticalement (scroll), la longueur du texte des questions et le nombre de modalités de réponse doivent être limités, notamment pour les batteries de questions. En effet, étant donnée la taille de l'écran, les batteries de questions doivent être limitées à cinq modalités de réponse. Cette limite s'est avérée particulièrement forte dans le cadre du projet comparatif avec les autres panels européens. Pour conserver la présentation sous forme de batterie de certaines questions des European Election Studies, le nombre de modalités de réponses a dû être limité à 6 pour ELIPSS alors que le GIP et le LISS panel ont pu utiliser les versions originales offrant 11 modalités de réponse. L'usage des batteries de questions est discuté dans la littérature, en particulier pour les web surveys (Couper, 2008). Comme le suggérait Mick Couper lors de sa venue à Paris en octobre 2013, le panel ELIPSS offre une occasion unique d'étudier la façon dont les gens répondent à des batteries de questions sur un écran de taille intermédiaire entre celle de l'ordinateur et celle du smartphone. Lors de l'enquête de novembre 2013 sur le couple et la famille, nous avons pu mener une expérimentation en accord avec l'équipe de recherche : la moitié des panélistes a répondu à certaines questions présentées sous forme de batteries, l'autre moitié a répondu à ces questions présentées sous forme questions nominales (une par page).

La participation aux enquêtes

Maintenir la participation

Les relances des panélistes pour leur participation mensuelle aux enquêtes reposent sur plusieurs dispositifs. Après avoir été invités au panel le premier jour de terrain par un message sur la tablette (et par mail lorsque l'adresse électronique est disponible), des messages automatiques de relance sont envoyés sur les tablettes des non-répondants les vendredis à partir de la seconde semaine de terrain. Les graphiques ci-dessous montrent l'évolution du nombre de répondants par jour à l'enquête de mai 2013 (première enquête avec des relances systématisées) et celle de novembre 2013. On constate une première vague de réponse après l'invitation pour les deux enquêtes, puis pour la seconde enquête, des pics plus visibles les week-ends. Dans ce dernier cas, il est difficile de savoir pour l'instant si l'on peut associer ce phénomène à un effet des relances qui ont lieu le vendredi.

Figure 5: Participation quotidienne aux enquêtes de mai et novembre 2013



La participation aux enquêtes dépend également des contacts réguliers des panélistes avec les gestionnaires du panel. Les premiers mois, les contacts consistaient principalement à traiter des problèmes techniques liés aux tablettes et la connexion 3G, des cas de casse, de perte et de vol de tablettes. Ceci a permis de définir les différentes procédures pour gérer ces situations. Des procédures de relances ont été élaborées pour suivre les panélistes ne répondant pas à plusieurs enquêtes. Ainsi, les « somnolents » (non-répondants à l'enquête en cours et à l'enquête précédente) et les « invisibles » (non répondants à au moins deux enquêtes) font l'objet de relances personnalisées par téléphone et par courrier postal chaque semaine.

Les non-répondants à l'enquête en cours ou les répondants l'ayant commencée sans la terminer[20] sont relancés en fonction de la charge de travail des gestionnaires de panel et des taux de non réponse de ces deux groupes.

[20] Les panélistes qui ont commencé le questionnaire bénéficient d'un délai supplémentaire d'un mois pour le terminer.

Les raisons de la non-réponse des panélistes évoquées lors de ces relances personnalisées sont souvent des problèmes techniques ou des cas de vols ou pertes de tablette. Dans le cas où il s'avère que les panélistes ne peuvent techniquement pas répondre aux enquêtes, ils sont suspendus du panel.

Afin de faciliter le suivi des panélistes, un outil en ligne de gestion du panel à été développé par le développeur de l'équipe ELIPSS[21]. Tous les échanges avec les panélistes sont ainsi répertoriés au quotidien par les gestionnaires du panel.

La participation mensuelle et l'attrition

Le tableau 7 décrit la participation aux enquêtes administrées depuis l'inclusion dans le panel. Jusqu'en novembre 2013, le taux de participation (ce taux est le COMR décrit par Callegaro, Disogra, 2008) est au-dessus de 85% et commence ensuite à décliner pour atteindre un taux de 77% à l'enquête de février 2014[22]. Si l'on constate une chute importante de la participation à cette enquête, il convient de noter que cette enquête est en réalité la seule à avoir connu un terrain de quatre semaines, sans délai supplémentaire et avec un problème technique qui a empêché l'envoi de mails de relance. L'enquête de mars 2014 connaît l'un des taux de réponse les plus élevés avec 90%, ceci après une campagne de relances personnalisées et un terrain rallongé de quelques jours. Des efforts importants ont en effet été mis en œuvre car il s'agit de l'enquête annuelle dans laquelle les caractéristiques socio-démographiques des panélistes sont mises à jour.

Le taux d'attrition commence à augmenter à la fin 2013 et atteint 8% en mars 2014. Ceci correspond aux premières désactivations des panélistes « invisibles » qui ont eu lieu à la suite de la systématisation des relances auprès de ce groupe.

Tableau 7: Liste des enquêtes administrées au panel ELIPSS depuis décembre 2012

Terrain	Enquête	Taux de réponse	Nombre d'invités	Attrition
déc.2012-mars 2013	Enquêtes et internet	99%	943	0%
avril 2013	Enquête annuelle 2013 (module signalétique)	91%	1012	0%
mai 2013	Enquête annuelle 2013 (module barométrique)	87%	1011	0%
juin 2013	Pratiques culturelles, médias et technologies de l'information	88%	1011	0%
juil.2013-août 2013	Fécondité, contraception, dysfonctions sexuelles	87%	1005	2%

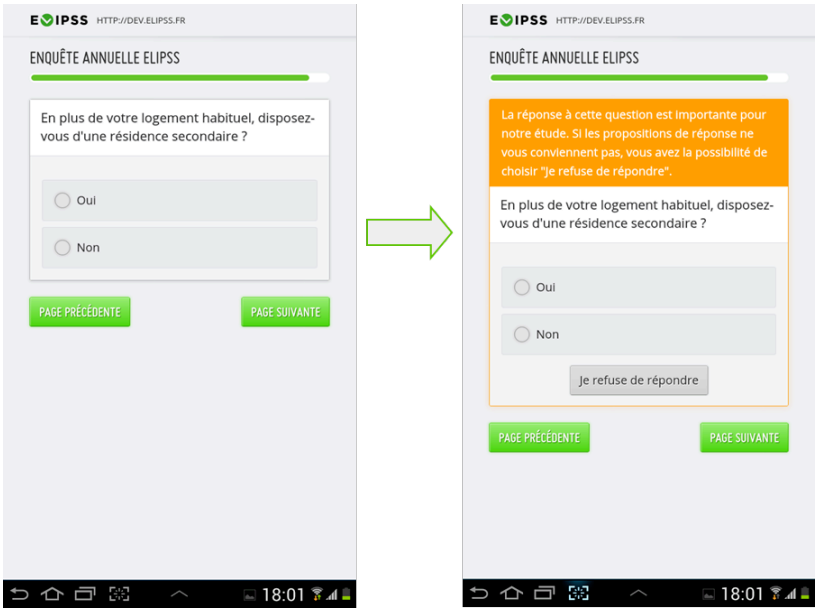
[21] Cet outil a bénéficié de l'expérience de l'institut de recherche néerlandais CentERdata, responsable du LISS panel, qui a mis à notre disposition leur propre système de gestion des panélistes.

[22] La vague pré-municipale de Dynamob de mars 2014 connaît un taux encore plus bas de 64% ce qui s'explique surtout par un terrain de 2 semaines seulement, ce qui était inédit pour les panélistes, comme pour l'équipe ELIPSS.

Terrain	Enquête	Taux de réponse	Nombre d'invités	Attrition
septembre 2013	Dynamique de mobilisation - vague 1	87%	996	2%
octobre 2013	Enquête sur les valeurs, l'environnement et l'énergie	85%	997	3%
novembre 2013	Situation de couple, intentions de fécondité et opinions sur la famille	88%	983	3%
déc.2013-janv.2014	Santé, travail et environnement. Enquête sur les expositions aux poussières inorganiques / Dynamique de mobilisation - vague 2	83%	993	3%
février 2014	Les relations entre générations au prisme des normes de solidarité et de justice sociale	77%	984	4%
mars 2014 (2 semaines)	Dynamique de mobilisation - vague 3 (pré-municipales)	64%	945	8%
mars 2014	Enquête annuelle 2014	90%	945	8%

La non-réponse partielle

La non-réponse partielle aux enquêtes fait également l'objet d'une attention particulière dans le design des questionnaires. Plusieurs options de non-réponse sont disponibles et sont choisies pour chaque question lors de l'élaboration des questionnaires. Ainsi, la non-réponse peut-être proposée comme une catégorie de réponse substantielle ou peut être distincte par l'affichage de boutons spécifiques (refus et/ou ne sait pas). Un message de relance peut être affiché après que le panéliste a essayé de passer à la question suivante et peut être accompagné ou non de boutons spécifiques de non-réponse (cf. copie d'écran ci-dessous).



La diffusion des données

L’accès aux données

Le Centre de données socio-politiques, qui est l'un des trois centres de données français en sciences sociales, est responsable de la documentation et de la diffusion des données produites dans le cadre du panel. Les enquêtes sont documentées selon la norme internationale Data Documentation Initiative et diffusées par internet.

Une fois la période d’exclusivité d’un an pour les équipes co-productrices des enquêtes, celles-ci seront répertoriées dans le catalogue d'enquêtes du portail du Réseau Quetelet (réseau français des centres de données en sciences sociales). A partir du quatrième trimestre 2014, les premiers fichiers de données seront accessibles gratuitement pour les chercheurs français et étrangers, les doctorants, les post-doctorants et les étudiants de master dans le cadre exclusif d’un projet de recherche.

Pour obtenir les fichiers de données, les demandeurs doivent signer une convention d’utilisation dans laquelle il s’engage notamment à respecter la confidentialité des répondants, à ne pas rediffuser des données à un tiers et à citer la source des données dans les publications.

L'appariement

Pour garantir la confidentialité des données, il sera impossible d'apparier toutes les données individuelles issues du panel entre elles. Seul le module module signalétique de l'enquête annuelle ELIPSS sera apparié de manière systématique à chaque fichier d'enquête. Toute demande d'appariement de données provenant de plusieurs enquêtes (en dehors des enquêtes longitudinales) est strictement encadrée et est soumise à l'examen du comité scientifique et technique.

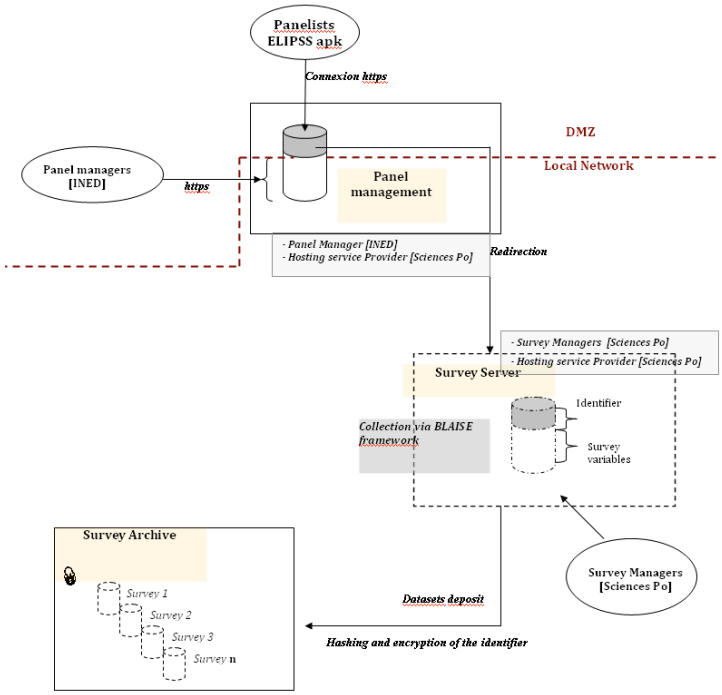
Par ailleurs, l'appariement des informations du panel avec des données extérieures (fiscales, santé, etc.) est exclu.

La confidentialité des données

Le panel ELIPSS a fait l'objet d'une déclaration à la Commission nationale de l'informatique et des libertés (CNIL) et est inscrit au registre du CNRS sous le numéro 2-12030. Une demande d'autorisation à la CNIL peut être nécessaire pour les enquêtes retenues par le comité scientifique et technique dès lors qu'elles comportent une ou plusieurs questions sensibles (origine ethnique, opinions politiques, philosophiques ou religieuses, appartenance syndicale, santé ou vie sexuelle).

La sécurité de l'information collectée à partir du panel ELIPSS est primordiale. Les données nominatives et les données d'enquêtes sont stockées dans deux systèmes d'information différents. D'un point de vue technique, le cryptage des données et les restrictions mises en place pour contrôler l'appariement des données garantissent également la confidentialité.

Figure 6 - Infrastructure informatique



Bilan et perspectives

La construction d'un dispositif tel que le panel ELIPSS recouvre des activités très diverses qui correspondent aux différentes étapes du cycle de vie des données, allant du projet d'enquête à la mise à disposition des données produites. En particulier, le pilote, destiné à tester la faisabilité d'un tel panel en France, a nécessité d'inventer l'inexistant. En premier lieu, le recrutement des panélistes a fait l'objet d'une procédure relativement complexe afin de déterminer la meilleure approche pour le développement du panel en 2015. Par ailleurs, l'originalité d'ELIPSS de fournir une tablette tactile et un abonnement 3G à chaque panéliste a nécessité la contractualisation avec un opérateur de téléphonie mobile, ce qui impliquait des aspects techniques, opérationnels, juridiques et financiers. Aussi, il a fallu développer en interne les outils informatiques puisqu'aucun outil existant ne remplissait l'ensemble des besoins du projet. Plusieurs de ces opérations, notamment le contrat avec l'opérateur de téléphonie, étaient des préalables indispensables à la construction du panel et sont maintenant définis pour le passage à 5000 panélistes.

Le pilote du panel ELIPSS a mis au jour des difficultés à prendre en compte et des pistes d'amélioration en vue du développement du panel en 2015. En premier lieu, l'utilisation d'une base d'adresses pour recruter les panélistes présente des inconvénients. Telle que fournie par l'INSEE, elle nécessite un important travail de nettoyage et de vérification des adresses. Elle implique aussi de décrire préalablement tous les membres du logement avant de sélectionner une personne pour l'inviter à participer au panel. Si ces contraintes sont communes à toutes les enquêtes menées à partir d'un échantillon d'adresses, le recrutement du pilote a rencontré des difficultés spécifiques à la procédure retenue et au timing du projet : le terrain qui s'est en partie déroulé pendant l'été ainsi que le retard pris du fait de la contractualisation avec l'opérateur ; l'invitation par courrier et les relances par téléphone qui ne se sont pas avérées aussi efficaces qu'escompté ; la signature de la convention de participation qui ajoute une

étape pour convaincre les personnes sélectionnées et formalise l'engagement à participer. Si la couverture réseau a été meilleure qu'attendue (aucune personne éligible n'a été exclue pour cette raison), la tablette a eu l'effet prévu en incitant les personnes à participer. Elle a aussi l'avantage de fournir un seul et même outil aux panélistes pour répondre aux questionnaires et dont l'aspect visuel est contrôlé.

Dans ce bilan, il convient également de souligner le rythme de travail soutenu que suppose la production d'une interrogation de 30 minutes par mois, en plus des activités de gestion de panel, de gestion de stock, des développements informatiques et de l'organisation du service auprès de la communauté académique.

En conclusion, il reste à tirer les enseignements du pilote pour continuer à construire et faire vivre le panel ELIPSS avec 5 000 individus jusqu'en octobre 2017.

Nous avons testé différents modes de contact, notamment par souci financier, mais force est de constater que le face-à-face est la stratégie la plus efficace pour recruter des panélistes. Il semble donc difficile de faire l'économie du recours à un institut de sondage pour le recrutement prévu au début de l'année 2015. Il en va de la taille et de la qualité de l'échantillon. Tout au plus, pouvons-nous envisager un protocole en deux phases : une lettre proposant aux personnes intéressées de s'inscrire directement sur le site web dédié et annonçant sinon la visite d'un enquêteur, puis un recrutement en face-à-face des non-répondants. En effet, la possibilité de s'inscrire par internet avait permis de recueillir facilement presque 15% des réponses positives au niveau du ménage. Pour des raisons méthodologiques et techniques, nous comptons continuer à équiper les panélistes de tablettes ayant toutes le même format, fonctionnant avec le même système d'exploitation et sur lesquelles l'application ELIPSS sera pré-installée. Cependant, ces choix en matière de recrutement et d'équipement internet ne sont pas financés actuellement et ont fait l'objet d'une demande de financement complémentaire auprès de la région Ile de France.

Par ailleurs, le recrutement des nouveaux panélistes en 2015 mérite d'être mieux préparé. Nous bénéficierons évidemment de l'expérience acquise lors du pilote. La procédure sera également plus simple et certainement entièrement prise en charge par l'institut de sondage. De plus, nous avons prévu de renforcer l'équipe en embauchant une personne chargée précisément de préparer et de suivre le terrain en face-à-face.

En dernier lieu, cette période sera une occasion unique de mener des expérimentations et d'étudier par exemple l'éventuelle professionnalisation en comparant les comportements de réponse des anciens et des nouveaux panélistes. Pendant le pilote, les travaux méthodologiques ont dû généralement passer au second plan par rapport à la mise en place du panel et à la production des enquêtes mensuelles car le projet avait largement été sous-dimensionné en termes de force de travail. Compte tenu du renforcement prévu de l'équipe, la recherche méthodologique, indispensable pour suivre et maintenir la qualité du dispositif, devrait occuper une plus grande place dans le projet.

Bibliographie

Blom A., Gathmann C., Krieger U., The German Internet Panel: Method and Results, 2013

Callegaro M., 2010 – Do You Know Which Device Your Respondent Has Used to Take Your Online Survey?, *Survey Pratice*, vol.3, n°6, (http://www.surveypractice.org/index.php/SurveyPractice/article/view/250/html)

Callegaro M., DiSogra C., 2008 – « Computing Response Metrics for Online Panels », *Public Opinion Quaterly*, vol.75, n°5, p.1008-1032

Couper M., 2008 – *Designing effective web surveys*, New York: Cambridge University Press

Das M., Ester P., Kaczmirek L.(eds.), 2011 – *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies*, Boca Raton: Taylor & Francis

Gombault V., 2013 – « L'internet de plus en plus prisé, l'internaute de plus en plus mobile », *INSEE première*, n°1452, juin

Knoef M., de Vos K., 2009 – The representativeness of the LISS, an online probability panel, Tilburg, CentERdata, 29p. (http://www.lisspanel.nl/assets/uploaded/representativeness_LISS_panel.pdf)

Leenheer J., Scherpenzeel A.C.,, 2013 – Does it pay off to include non-internet households in an internet panel?, *International Journal of Interent Science*, vol. 8, n°1, p.17-29

Scherpenzeel A., 2009 – Start of the LISS panel: Sample and recruitement of a probability-based internet panel, Tilburg, CentERdata, 9 p. (http://www.lissdata.nl/assets/uploaded/Sample_and_Recruitment.pdf)

DIME Web

Présentation de l'instrument

Dime Web aide les chercheurs à étudier des traces numériques et en particulier le web par un ensemble de moyens que l'on regroupe sous le nom de méthodes numériques. Les chercheurs en sciences humaines ont l'opportunité d'accéder à des données que l'on peut considérer comme une nouvelle incarnation du social, mais qui sont difficiles à exploiter en raison de leur taille, de leur dynamicité ou de leur complexité. Dime Web procure outils, formation et accompagnement méthodologique aux chercheurs qui ont besoin d'aide pour inclure des terrains numériques dans le périmètre de leur recherche.

L'équipe est composée de deux ingénieurs ayant une expérience de la recherche. Mathieu Jacomy est responsable du projet, en effectue la valorisation et développe les interfaces des outils. Benjamin Ooghe-Tabanou développe les traitements et stockages sur serveur et les scripts de captation des traces numériques. Tous deux accompagnent les chercheurs selon les projets. Ils bénéficient des ressources du médialab pour compléter leurs compétences (designers, développeurs et chercheurs).

Enjeux

Dime Web a vocation à procurer aux chercheurs des *moyens adaptés*, ce qui comprend la mise à disposition d'outils aussi bien que la transmission des compétences nécessaires à leur utilisation.

Les outils disponibles ne sont pas toujours adaptés aux chercheurs en sciences humaines. La plupart des logiciels permettant de traiter des traces numériques ont été développés pour d'autres usages (veille, gestion…) et selon d'autres critères de qualité (performance, rentabilité…). Les chercheurs ont une longue pratique du détournement de ces outils, non sans inconvénients majeurs. La première difficulté est de traduire des objets de recherche dans le fonctionnement de l'outil (et vice-versa). La deuxième difficulté est d'accéder à la documentation des traitements effectués (algorithmes, conversions et approximations), souvent inexistante. Enfin la dernière difficulté, la plus courante, est la nécessité d'une expertise technique élevée (lignes de commandes, expressions régulières…). Dime Web aide les chercheurs à affronter ces problèmes, que ce soit par l'accompagnement, la documentation ou la création de nouveaux outils.

Le coût des solutions logicielles est un autre frein à la démocratisation des méthodes numériques dans les sciences sociales. Leurs chercheurs ne constituent pas un marché économique stratégique pour les sociétés éditrices de logiciel, avec pour conséquence l'accroissement d'un retard technologique dans ces disciplines. Il est cependant possible de débloquer l'accès à de nouveaux moyens lorsque les technologies existent et que ce sont les interfaces et la documentation qui manquent. C'est particulièrement le cas de la collecte de données formatées sur le web, appelée "scraping": il est possible de déployer ce type de technique en quelques lignes de code, mais aucune interface graphique n'était disponible jusqu'à très récemment (début 2014). Dans de tels cas il existe une opportunité économique, puisque l'effort nécessaire est faible en comparaison des services rendus. Dime Web participe à la démocratisation des méthodes numériques en

développant des briques logicielles qui manquent, lorsque le rapport coût/bénéfice est favorable. Nos développements sont toujours mutualisés, que ce soit en déployant des outils en ligne accessibles à tous ou en publiant le code source sous licence libre. Nos efforts s'ajoutent ainsi à ceux d'autres équipes qui partagent les mêmes objectifs, et avec qui nous nous coordonnons (notamment l'équipe de Richard Rogers à Amsterdam, Digital Methods Initiative et l'équipe de Paolo Ciuccarelli à Milan, Density Design Lab).

Fonctionnement

Les chercheurs bénéficient d'un accompagnement par projet. Un dossier est requis et la sélection est effectuée par un Comité Scientifique et Technique propre au volet Web, sur des critères de pertinence scientifique et de faisabilité technique. Chaque projet a ses propres thématiques de recherche, problématiques, méthodes et besoins. Nous essayons d'apporter une réponse personnalisée à chaque projet en nous adaptant aux exigences méthodologiques de chacun. Ce sont les besoins techniques qui varient en pratique, la nécessité d'un accompagnement et d'un transfert d'expertise étant constante. Le format de notre accompagnement est amené à se stabiliser sous la forme d'une offre construite au fur et à mesure que nous sortons de la phase de construction de l'équipement.

Parallèlement à l'appui aux chercheurs nous développons le crawler *Hyphe* qui permet de construire un corpus de documents web personnalisé. Son ambition est de lever un verrou méthodologique majeur dans l'étude des communautés en ligne. Une section ultérieure de ce document le présente plus en détail.

État d'avancement du projet

Aavancement des projets sélectionnés

OpenMarriage - Analyse dynamique des controverses sur le « mariage pour tous », avec Eric Dagiral (CERLIS). Appel à projets Automne 2012.

Une collecte de commentaires sur une selection d'articles dans LeMonde.fr, avec publication du code de scraping en licence libre.

Une réunion de lancement en mars 2013, une session pour collecter les commentaires d'articles Le Monde (2j, réutilisable), puis une réunion en décembre 2013 (en plus d'échanges en ligne). Des ennuis de santé ont ralenti notre collaboration, qui est actuellement en stand-by.

SitPol - Cartographie exploratoire du web politique, avec Dinah Galligo à la Bibliothèque de Sciences Po. Appel à projets Automne 2012.

Nous avons formé les documentalistes de la bibliothèque à l'utilisation de Hyphe, et nous avons crawlé leur liste qualifiée de sites web. Les fonctions de définition personnalisée d'entités web de Hyphe ont été largement mobilisées. Le réseau de liens hyphertextes a été produit, et il est valorisé dans un autre projet "sitothèque en ligne" (en cours et géré directement par la bibliothèque de Sciences Po).

Une réunion de lancement en mars 2013, une formation Hyphe fin avril 2013, une session de développements spécifiques de Hyphe de deux semaines en juillet 2013, suivi du lancement des crawls, et une réunion de clôture en octobre 2013.

Aesif Web - Les agences européennes de sécurité intérieure et de gestion des frontières vues du web, avec Didier Bigo, CERI. Appel à projets Automne 2012.

Nous avons construit une liste de 172 requêtes Google et nous avons collecté les 500 premiers résultats pour chacune. Nous avons fouillé le contenu textuel de ces pages (n-grammes) et construit différentes visualisations (cf. copie d'écran ci-dessous).

Un sprint inaugural de deux jours en avril 2013, deux semaines de développements spécifiques mi-avril et début août, une réunion et un sprint technique en octobre 2013. Réunion de clôture début 2014.

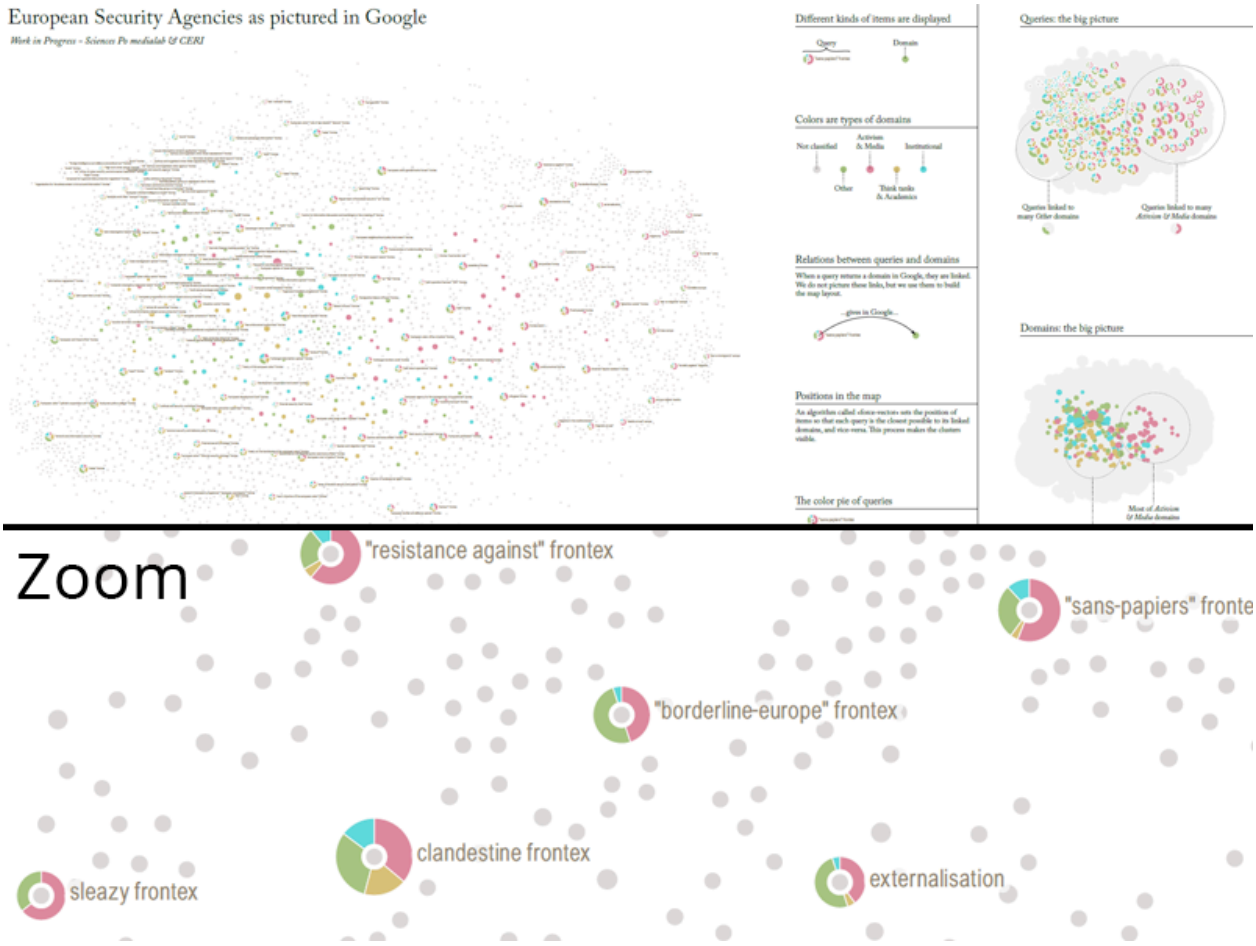


Figure 7 - Requêtes Google à propos de Frontex pour AESIF Web

Amour2Pré - L'amour est dans le pré et ailleurs, une sociologie de la croyance en l'amour à travers l'articulation de la vie conjugale et la réception des productions culturelles. Christophe Giraud, CERLIS. Appel à projets Printemps 2013.

Le projet est en attente de la diffusion de l'émission L'Amour est dans le Pré. Les tweets comportant une sélection de hashtags relatifs à l'émission sont archivés en continu (stream live et requêtes sur l'api), pour un total de 30,000 tweets pour l'instant.

Une réunion préparatoire a eu lieu fin 2013.

HateWeb - Représenter l'espace des mobilisations islamophobes en Italie : structures, évolutions et magistères intellectuels. Tommaso Vitale, CEE Sciences Po. Appel à projets Printemps 2013.

Le projet comporte essentiellement une collecte de sites web avec Hyphe et un traitement des contenus textuels. Hyphe est actuellement déployé, une formation des chercheurs a eu lieu et le corpus est en partie constitué (cf. copie d'écran de Hyphe ci-dessous). Une réunion de suivi a également eu lieu en avril 2014.

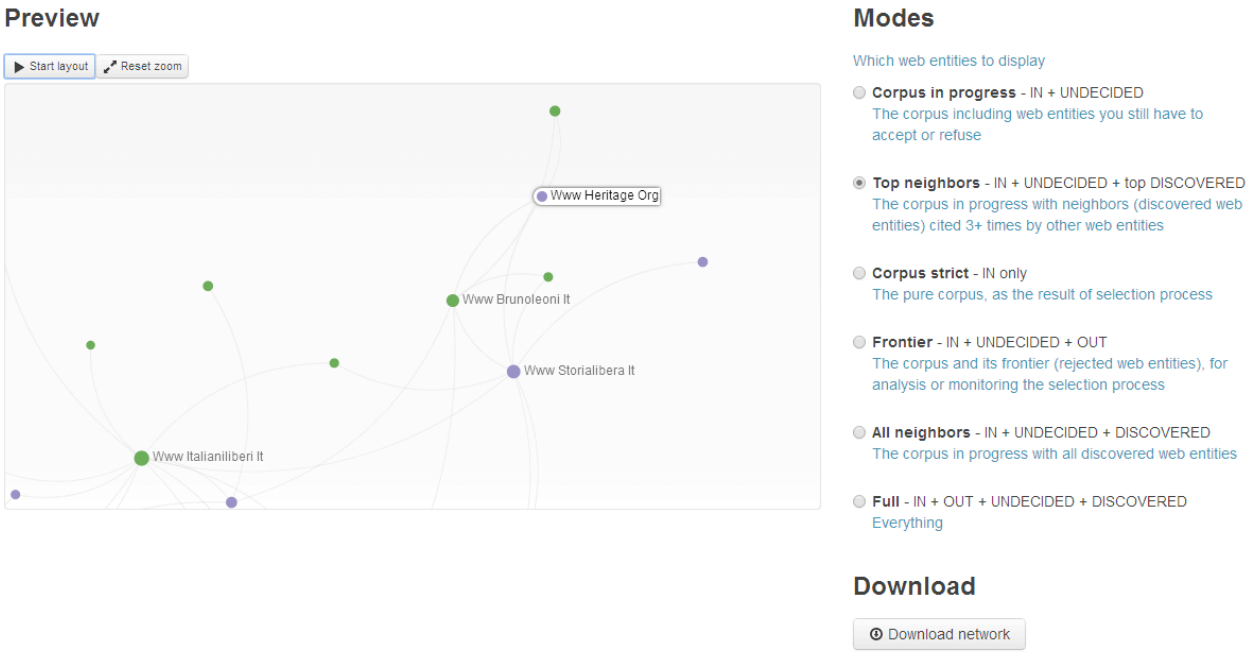


Figure 8- Corpus HateWeb en cours de constitution dans Hyphe

Projets à la marge - Nous avons mobilisé Hyphe à diverses occasions pour recueillir des retours d'utilisateurs, en dehors de l'appel à projet : pour une cartographie des sites de l'institution Sciences Po avec Ève Demazières, pour une analyse de la controverse sur le planning familial puis sur la césarienne, dans le cadre d'un projet du médialab pour l'OMS, pour l'analyse de la controverses sur l'adaptation au changement climatique, et pour des projets d'étudiants.

Autres avancements opérationnels

Développement de Hyphe - Deux versions publiées en 2013, facilement installables sur différentes distributions Linux. Contribution à la création du paquet générique de ScrapyD pour CentOS. Dans l'interface, intégration de Domino.js, interface de diagnostic CSV, export de graphes et "corpus overview" réalisé à 50%. Hyphe est décrit en annexe 4 p. 57.

Autres développements - ScienceScape, outil pour avoir des graphes facilement avec des données scientométriques. Utilisé par des étudiants et lors de formations à Gephi pour obtenir des données faciles à interpréter.

Dissémination - Nous avons donné des présentations sur Dime Web, Hyphe et/ou les méthodes numériques à 11 occasions en 2013 (King's College à Londres, Université d'Amsterdam, Université de Göteborg, CEREQ à Marseille…).

Vie du projet - Nos tâches courantes ont comporté notamment la rédaction de deux appels d'offres, la mise en place d'une feuille de route, l'organisation de deux CST, la mise en place du reporting du temps de travail, et la proposition d'une offre produit plus détaillée. Nous avons également tenu une série d'entretiens avec divers chercheurs intéressés par le projet (Yann Algan, Fabrice Epelboin, Dana Diminescu, Lionel Villard, Tommaso Vitale, Cécile Brousse…)

Publications académiques - L'écriture d'un article est prévu dans le projet AESIF Web. L'équipe opérationnelle a soumis un article scientifique à la revue PlosOne au sujet d'un algorithme de spatialisation de réseaux que nous utilisons dans Hyphe et Gephi. Une première révision de l'article a été re-soumise en mars 2014. Une publication présentant les enjeux du logiciel Hyphe est en corus de rédaction.

Perspectives

Format de l'offre - Le format en appel à projets deux fois par an s'avère insatisfaisant pour plusieurs raisons. Les chercheurs qui ont besoin de notre soutien ont souvent une vision floue des possibilités offertes par les méthodes numériques, ainsi que de leurs contraintes et limitations: constituer le dossier est donc une tâche difficile à réaliser, voire dissuasive. Trois des neuf dossiers que nous avons reçus en deux appels étaient si insuffisamment motivés que le CST n'a tout simplement pas pu les évaluer. En outre une partie des demandes informelles que nous recevons prennent la forme d'un soutien ponctuel et rapide, que le calendrier des appels à projets empêche de satisfaire parce qu'il est trop lent, et que le dossier est trop lourd à constituer au regard de la taille du soutien demandé. Nous sommes donc en train de faire évoluer l'offre pour prendre en compte différemment les petits et gros projets, avec la perspective d'inciter les chercheurs à découvrir les méthodes numériques sur des petit projets et proposer des montages plus ambitieux dans un second temps.

Plan d'affaires - La contrainte d'auto-financement qui constitue l'une des règle de l'Equipex pose encore de nombreuses questions. Pour gérer cet aspect nous sommes également en train d'écrire un plan d'affaires, comprenant notamment la réflexion précédente sur le format de l'offre. Le mode de financement de Dime Web pourrait se développer par une offre de services payants: accompagnement méthodologique, hébergement de corpus sur infrastructure dédiée, formation, développement informatique spécifique… La phase de test nous permet aujourd'hui d'envisager cette offre au vue des besoins exprimés jusqu'ici. Si l'offre nous semble rencontrer son public, il reste de nombreuses questions: prix et équilibre budgétaire, volume des demandes, modes de contractualisassions, forme juridique de l'entité gérant l'offre de service...

Ces questions constituent notre principal défi pour les années à venir.

Règles d'éthique

Nos échanges avec le Contact Informatique et Libertés de Sciences Po, nous ont permis d'identifier la bonne procédure pour la protection des données personnelles. L'équipe Dime Web s'engage à veiller à respecter les règles de protection des données

personnelles (déclaration de la base et anonymisation si nécessaire). Dans la très grande majorité des cas, l'activité de Dime Web consiste à récolter des données web publiques sans créer de bases de données de noms ou profils même si ceux-ci peuvent être présents dans les données. Ainsi la grande majorité de l'activité ne constitue pas d'enjeu d'éthique.

L'équipe s'engage également à alerter et conseiller les chercheurs avec lesquels elle collabore à ces enjeux.

Nous respectons par ailleurs les restrictions afférentes aux données collectées chez des prestataires privés, typiquement l'interdiction que fait Twitter de rediffuser les données brutes.

Enfin, nous publions le code de nos logiciels sous license libre et open source, essentiellement sur le compte GitHub du médialab.

Annexes

Annexe 1 - Extrait de l'accord de consortium

6.4 LE CONSEIL SCIENTIFIQUE

Cette instance consultative est informée de l'utilisation de l'équipement d'excellence DIME-SHS par le COORDINA TEUR qui participe, sans voix délibérative, aux séances du CONSEIL SCIENTIFIQUE.

6.4.1 Composition du CONSEIL SCIENTIFIQUE

Le CONSEIL SCIENTIFIQUE est composé de 12 personnalités expertes et reconnues à l'international dans le domaine des méthodes en sciences sociales, nommées par le COMITE DE PILOTAGE. Au moins la moitié de ces personnalités sont des experts de leur domaine extérieurs aux PARTERNAIRES.

Les membres du CONSEIL SCIENTIFIQUE siègent en qualité d'experts indépendants et ne représentent en aucun cas la ou les institutions auxquelles ils appartiennent, que ce soit à titre professionnel ou non.

Le CONSEIL SCIENTIFIQUE est garant d'une représentation équilibrée des trois instruments de DIME-SHS.

Le CONSEIL SCIENTIFIQUE compte 3 personnes spécialistes des questions d'éthique.

Le Président du CONSEIL SCIENTIFIQUE est désigné par le COMITE DE PILOTAGE et a en charge la convocation des réunions du CONSEIL SCIENTIFIQUE, la rédaction des comptes-rendus, et leur diffusion auprès des membres du Conseil scientifique, du COMITE DE PILOTAGE et du COORDINATEUR.

Afin d'assurer la mission qui lui est confiée dans le cadre du PROJET, le COORDINATEUR est invité permanent du CONSEIL SCIENTIFIQUE et peut inscrire des points à l'ordre du jour des réunions. Il est destinataire des comptes rendus, procès-verbaux et documents en tous genres produits par le CONSEIL SCIENTIFIQUE.

Le COMITE DE PILOTAGE décide de la durée des mandats (3 ou 4 ans par exemple) et des modalités de renouvellement des membres du CONSEIL SCIENTIFIQUE.

6.4.2 Réunions du CONSEIL SCIENTIFIQUE

Le CONSEIL SCIENTIFIQUE se réunit au moins une fois par an, sur convocation de son président. Le recours à des dispositifs collaboratifs (téléconférence, visio-conférence) sera possible. Des réunions extraordinaires peuvent être organisées par le président du CONSEIL SCIENTIFIQUE, en cas d'urgence notamment, sur demande écrite et motivée du COORDINATEUR, d'un ou plusieurs PARTENAIRES, ou membres du CONSEIL SCIENTIFIQUE.

Sauf urgence, le président adresse l'ordre du jour aux membres du CONSEIL

SCIENTIFIQUE au moins quinze (15) jours avant la réunion.

6.4.3 Règles de décision au sein du CONSEIL SCIENTIFIQUE

Le CONSEIL SCIENTIFIQUE est valablement réuni si les trois quarts (3/4) de ses membres sont présents ou représentés. Si lors d'une réunion le quorum n'est pas atteint, le CONSEIL SCIENTIFIQUE est convoqué une seconde fois, dans un délai qui ne peut excéder 4 semaines à compter de la date de la réunion initiale. A la suite de cette seconde convocation, le CONSEIL SCIENTIFIQUE est valablement réuni si le 1/4 de ses membres est présent ou représenté.

Les membres du CONSEIL SCIENTIFIQUE peuvent recevoir, pour une réunion donnée, un mandat de représentation d'un autre membre, dans la limite d'un mandat par réunion. Tous les membres du CONSEIL SCIENTIFIQUE disposent d'une voix.

6.4.4 Rôle du CONSEIL SCIENTIFIQUE

Le CONSEIL SCIENTIFIQUE est notamment chargé :

- de formuler des orientations scientifiques pour le COORDINATEUR et le COMITE DE PILOTAGE et le cas échéant de faire des propositions de modification du projet scientifique au COORDINATEUR et au COMITE DE PILOTAGE ;
- d'assurer une prospective scientifique sur les besoins à venir en matière de données pour les projets et de proposer des priorités dans la mobilisation des bases de données et les appariements ;
- de donner un avis quant au fonctionnement de l'équipement d'excellence DIME-SHS du point de vue des utilisateurs français comme étrangers ;
- d'assurer une veille technologique, méthodologique, juridique et éthique sur l'accès aux données confidentielles en lien avec les développements internationaux ;
- de faire des propositions quant à l'animation scientifique de l'équipement d'excellence DIME- SHS ;
- de maintenir une veille sur les questions d'anonymisation/anonymat, avec l'aide ponctuelle d'experts extérieurs et/ou du comité de concertation pour les données en sciences humaines et sociales (CCDSHS) ;
- d'assurer une veille et de faire des propositions quant aux partenariats avec d'autres centres fournisseurs d'accès au plan national afin de favoriser l'harmonisation des procédures et des standards ainsi qu'une bonne synergie ;
- de veiller à la présence de l'équipement d'excellence DIME-SHS au sein des projets et infrastructures en développements au plan européen et international.

Annexe 2 - Contours documentaires d'une enquête qualitative

1. Documents indispensables – ou presque	2. Documents complémentaires	3. Documents pour les archives
Projet(s) de recherche	Parties des rapports d'évaluation des chercheurs relatifs au projet	Notes manuscrites du chercheur préparatoires au projet
Demandes de financement obtenues	Demandes de financement échouées	
Budget global	Etat final des dépenses liées au projet	Factures, ordres de mission, états de frais
Correspondance papier entre les membres de l'équipe de recherche (et dans le cas des thèses, avec le directeur de thèse)	Sélection éventuelle de correspondance électronique entre les membres de l'équipe ou avec le directeur de thèse	Documents correspondants non-sélectionnés
Correspondance papier avec les enquêtés / le « terrain »	Sélection éventuelle de correspondance électronique relative au terrain	Documents correspondants non-sélectionnés
Compte-rendu ou procès-verbaux de réunions d'équipe	*Lorsque les PV sont indisponibles* : notes manuscrites ou électroniques de réunions d'équipe	*Sinon* : notes manuscrites ou électroniques de réunions d'équipe
Bibliographie préparatoire de l'enquête	Littérature grise préparatoire à l'enquêtes – si les documents ne sont pas disponibles ailleurs	Littérature grise et revue de presse éventuelle de l'enquête quand les documents sont disponibles ailleurs – surtout si annotés par le chercheur
Principes de sélection des interviewés, courriers et annonces utilisés pour le recrutement	Tous autres documents relatifs au recrutement des interviewés (le cas échéant)	

Tous documents définitifs organisant la réalisation du terrain : grille d'observation, d'entretien, matériel projectif, questionnaires, etc.	Documents préparatoires aux documents de terrain : versions intermédiaires et tests	Notes relatives à ces documents
Collecte : carnets de notes, enregistrements, transcriptions, corpus documentaire	Doubles annotés de certaines transcriptions ou documents	
Tous documents relatifs à la démarche d'analyse, à sa conception et/ou sa mise en œuvre	Des extraits ou exemples de mise en œuvre du travail d'analyse	Notes manuscrites et correspondance
Production scientifique non disponible : contributions à des colloques, articles non publiés…	Pre-prints des articles disponibles seulement de façon payante	Brouillons, versions préparatoires, tirés-à-part.

Annexe 3 - DIME-Quanti

Calendrier

<table>
<tr><td></td><td></td><td>Recrutement des panélistes</td><td>Opérateur téléphonique</td><td>Outils informatiques</td><td>Enquêtes</td></tr>
<tr><td rowspan="10">2012</td><td>mars-12</td><td rowspan="2">Tirage de l'échantillon par l'INSEE</td><td></td><td>Début développement des applications panélistes, gestion du panel et des enquêtes</td><td></td></tr>
<tr><td>avr-12</td><td>Appel d'offres opérateurs</td><td></td><td></td></tr>
<tr><td>mai-12</td><td>Préparation des documents panélistes</td><td></td><td>Ouverture du site elipss.fr et de l'application de recrutement</td><td>Formation Blaise</td></tr>
<tr><td>juin-12</td><td>Envoi des courriers d'invitation et de relance
Appel d'offres instituts de sondage</td><td>Choix de Bouygues Télécom</td><td></td><td rowspan="2">Groupe de travail
Enquête annuelle ELIPSS</td></tr>
<tr><td>juil-12</td><td>Choix de TNS Sofres
Relances téléphoniques</td><td rowspan="3">Négociation du contrat</td><td></td></tr>
<tr><td>août-12</td><td></td><td></td><td></td></tr>
<tr><td>sept-12</td><td>Préparation du terrain par TNS Sofres</td><td>1e version de l'application panélistes</td><td>1e CST DIME Quanti</td></tr>
<tr><td>oct-12</td><td rowspan="5">Envoi des conventions de participation
Recrutement par TNS SOFRES</td><td>Signature du contrat</td><td>1e version de l'application gestion du panel
et des enquêtes</td><td>Appel à projets 2012</td></tr>
<tr><td>nov-12</td><td rowspan="4">Envoi des tablettes
Activation des abonnements
Formation téléphonique</td><td>Programmation, design et tests de la 1e enquête</td><td></td></tr>
<tr><td>déc-12</td><td></td><td rowspan="4">1e enquête Pratiques numériques</td></tr>
<tr><td rowspan="3">2013</td><td>janv-13</td><td></td></tr>
<tr><td>févr-13</td><td></td></tr>
<tr><td>mars-13</td><td></td><td></td><td></td></tr>
</table>

Étapes-clés 2014-2017

		Recrutement des panélistes	Opérateur téléphonique	Outils informatiques	Enquêtes
2014	janv-14				
	févr-14				
	mars-14			Amélioration de l'outil de gestion de panel	
	avr-14				
	mai-14	Evaluation du pilote Décision quant à la poursuite du projet		Développement d'un outil de gestion de stock	Appel à projets 2014 ouvert à toute la communauté académique
	juin-14				
	juil-14				
	août-14				
	sept-14	Appel d'offres instituts de sondage			
	oct-14				
	nov-14	Recrutement des nouveaux panélistes			
	déc-14				
2015	janv-15		Envoi des tablettes Activation des abonnements		
	févr-15				
	mars-15				
	avril-15				
2016					
2017	oct-17		Fin du contrat		

Annexe 4 - Présentation du crawler Hyphe

Les chercheurs en sciences humaines et sociales utilisent le web comme source d'information, de connaissance et terrain d'échange social. Ils ont besoin de collecter des traces d'activités pour nourrir leurs hypothèses de recherche et pour cela cherchent à construire des corpus documentaires. Or, les unités documentaires peuvent prendre plusieurs formes. La diversité des sources que l'on peut collecter (tweets, pages, sites, documents à télécharger) complique énormément la construction de corpus robustes et pose la question de l'unité de ces corpus.

Nous apportons une solution méthodologique à ce problème en développant le crawler Hyphe fondé sur l'idée de collecter des données via la définition de web entités fournies par le chercheur. Ces web entités précisent la granularité de chaque collecte et garantissent l'unité de chaque corpus. Cette idée nous démarque des autres crawlers et nous permet de faire le pont entre les concepts de la recherche (Acteurs, Arguments, Position, Influence, Réseaux, Diffusion) et la structure technique du web (URLs, …). Hyphe est un crawler construit pour équiper les problématiques de recherche.

Les autres crawlers (Heritrix, Issue Crawler…), utilisés par les chercheurs, sont en réalité détournés de leur fonction première: l'indexation d'informations. Ces outils ne sont pas faits pour construire des corpus, mais pour collecter des sites ou pages, et les indexer pour les rendre plus faciles à manipuler. Ces crawlers recopient et restituent l'arborescence technique des contenus web alors que les problématiques qui sous-tendent la construction de corpus à des fins de recherche sont tout autres. Le chercheur a besoin de retrouver ses objets de recherche, et ceux-ci ne s'organisent pas d'eux-mêmes selon le découpage technique du web. Par exemple, une personne a plusieurs blogs, tandis qu'un blog a plusieurs auteurs. Si le chercheur souhaite étudier les traces d'acteurs, il aura besoin de crawler différentes sources selon une granularité spécifique que nous appelons des "web entités".

Le principe de web entité est fondé sur une analyse morphologique des URLs. Hyphe est capable de comprendre la structure des URLs pour atteindre et différencier les contenus en exploitant leurs "termes" (un terme entre chaque slash "/"). Une web entité peut être vue comme une référence à un ou plusieurs des termes d'une ou plusieurs URLS. En définissant ses propres web entités, le chercheur décide non seulement de guider la collecte semi-automatique d'information mais il peut également les étiqueter. Les web entités traduisent en pratique les objets de recherche. Le chercheur choisit celles qui correspondent à ses questions de recherche. Il peut le faire a priori (au moment où la recherche est lancée) mais aussi et surtout a posteriori, après la collecte des pages. Utiliser Hyphe permet de garder la main sur la construction de son corpus de recherche et évite l'effet boite noire d'un grand nombre d'outils de collecte et de construction.

Nous illustrons maintenant notre méthode et l'utilisation de Hyphe en décrivant une problématique de recherche concrète. Un chercheur souhaite étudier la dynamique du mariage pour tous dans l'espace public, et notamment sur le web. Son point de départ consiste en une liste de blogs qu'il a préalablement identifiés. Il utilise Hyphe pour prospecter, en partant de l'hypothèse que ces blogs sont des points d'entrée vers d'autres ressources pertinentes. Comme la plupart des crawlers, Hyphe reconnaît les liens hypertextes et peut les suivre pour étendre le corpus (sur décision du chercheur). Notre chercheur découvre que ses blogs de départ pointent vers différentes choses: des blogs, des articles du Monde.fr, des sites d'institutions… En allant voir à la main le

contenu de ces ressources, il réalise que seuls certains de ces articles sont pertinents au regard de sa problématique de recherche (ils parlent du mariage pour tous). Il peut alors sélectionner ces articles et en faire de nouvelles web entités. Hyphe permet de déclarer n'importe quel ensemble de pages comme web entité, quel que soit son niveau de granularité : une page, un morceau de site, un nom de domaine, ou des combinaisons entre elles. Ainsi, ce processus ne nécessite pas de déclarer une web entité pour chaque article du Monde.fr : seuls ceux parlant du mariage pour tous deviennent des web entités. Le reste du site Le Monde.fr peut rester une seule web entité qui ne contiendra pas les articles extraits. Le chercheur construit son corpus web de sorte qu'il contient des acteurs (blogs, institutions) et des articles de presse (articles du Monde.fr). Il peut exporter ce corpus pour l'étudier dans d'autres outils, par exemple pour analyser quels acteurs citent quels articles.

Puisque la recherche est toujours un processus itératif dans lequel le chercheur est amené à redessiner en permanence le contour de ses objets d'étude, c'est donc souvent après la collecte que le chercheur souhaite redéfinir ses web entités. Les autres crawlers ne permettent en général pas de le faire, à moins de tout recommencer, depuis la collecte. Le blocage se trouve au niveau de l'étape d'indexation, une opération coûteuse en temps et en puissance de calcul qui permet de rendre les données collectées plus accessibles et calculables. Nous avons développé pour Hyphe un moteur d'indexation différent et original qui permet de redéfinir des web entités sans refaire l'indexation. Nous utilisons une stratégie différente des autres crawlers, qui permet la déclaration dynamique de chaque web entité, c'est-à-dire à tout moment de la méthode. Cette particularité permet de modifier le périmètre de chaque web entité au moment où cela fait sens pour le chercheur.

En conclusion, Hyphe est un outil développé spécifiquement pour les chercheurs en sciences sociales. Sa construction spécifique lui permet d'être plus adapté aux exigences méthodologiques des humanités numériques. Bien qu'il soit encore en développement, son code est accessible en ligne en tant que logiciel libre.