CAHIER DES CHARGES
version intermédiaire au 22 août 2020

PRÉSENTATION DE L'APPLICATION

Sondyn est un projet de sondages dynamiques. Le projet est porté par une société éponyme au capital ouvert à toute personne voulant participer au projet. À terme, le programme créé est laissé à la communauté. Il sera donc développé sous une licence libre qui reste à définir.

OBJECTIFS DE L'APPLICATION

1. Remontée de sondages

Le principe est de permettre à toute personne de proposer ses propres sondages en ligne auprès d'un cercle d'« amis » (dans le sens des réseaux sociaux : personnes avec laquelle on est en contact). Le résultat de ces sondages peut être compilé et remonté plus haut afin de fournir des résultats plus fiables (en appliquant alors les correctifs habituels) et de donner des valeurs significatives à des échelons souhaités, tels que département, région, voire niveau national.

2. Choix des questions et réponses posées

Chaque participant choisit les questions mais aussi les choix de réponses multiples à proposer, par ex. une liste de candidats potentiels. La remontée d'information est alors double : c'est l'existence d'un sondage avec certains choix qui est retransmise.

PORTÉE GÉOGRAPHIQUE DU PROJET

Bien que le code soit ouvert et utilisable dans le monde entier, ce sont essentiellement des résidents français qui sont visés.

CONCURRENCE

Ce projet concurrencera les instituts traditionnels de sondages sur le fond et sur la forme.

Sur le fond :
Les instituts classiques proposent des listes de candidats à des personnes qu'ils sondent. Pour nous, suivant les questions, et surtout s'il s'agit d'élections nominatives (ex : la présidentielle), ce sont les personnes interrogées qui vont donner le nom de leurs favoris, librement, sans restriction de liste pré-établie.
Cette manière est susceptible de remettre en cause une certaine crédibilité des instituts classiques qui, traditionnellement présentaient des listes toutes faites de candidats que leurs clients leurs donnaient, et limitaient implicitement le choix des sondés.

Sur la forme :
Hormis pour le lancement du projet qui nécessite des fonds et des investisseurs, à terme l'accès aux sondages sera rendu libre : il n'y aura plus de client qui paie. (En tout cas plus de client chez Sondyn, les instituts concurrents devraient continuer à avoir leurs clients et à faire payer leurs prestations.)

LA BASE ET LE NOMBRE D'UTILISATEURS

Un sondage traditionnel est effectué auprès d'un millier voire 3 000 personnes, tout en recueillant bon nombre de données supplémentaires telles que la localisation et les CSP en vue d'appliquer des coefficients et coller au plus vrai des résultats.

Ce ne sera pas notre cas (par choix, il n'est pas envisagé de recueillir trop de données personnelles). Alors nous devrons contre-balancer cette volonté par une base plus étendue de personnes sondées. 10 000 personnes pour un niveau national peut être envisagé.

Ce chiffre de 10 000 personnes minimum apportera de la crédibilité au projet dans la mesure où nous dépasserons le nombre de personnes sondées par les instituts concurrents ;

Une utilisation du programme par quelques centaines de personnes pourrait être considéré comme un échec.

Pour donner une autre fourchette de nombre d'utilisateurs potentiels : lors de questions anodines sur des réseaux sociaux, un petit sondage ordinaire peut recueillir 200 ou 500 votes. À l'inverse, au niveau national, à titre de repère, 1,5 million de personnes ont fait l'effort de s'inscrire devant leur ordinateur ou téléphone à l'occasion du référendum d'initiative populaire contre la privatisation des aéroports de Paris. Il s'agit de cas extrêmes.

Parmi les utilisateurs réels (en n'évoquant pas la très probable création de bots), il y aura les utilisateurs ponctuels (ceux qui s'inscrivent une fois et ne reviennent plus), les occasionnels (moins d'une dizaine de connexion dans l'année) et les actifs. Un utilisateur peut, par mégarde, ouvrir plusieurs comptes, sachant que seul le dernier compte actif comptera. Enfin, il y aura le cas des utilisateurs multi-actifs qui choisiront d'ouvrir plusieurs comptes pour orienter les sondages vers leurs idées. Tant pis, dans ce cas ce seront comme s'il y avait plusieurs utilisateurs enregistrés (de la même manière que des gens ne se connecteront jamais, nous pouvons nous dire que ceux-là se connectent pour plusieurs d'entre eux).

Lorsque nous évoquons le terme d'« utilisateur » sans autre précision, nous sous-entendrons les utilisateurs occasionnels et les actifs.

Nous pouvons maintenir le chiffre que nous venons d'annoncer de 10 000 utilisateurs, et l'encadrer par des cas possibles de 1 000 à 50 000 personnes, sachant que si ce dernier chiffre venait à être atteint ou dépassé, ce serait un succès, qui permettrait alors une éventuelle révision ou refonte du code.

sondyn.fr
Sondages dynamiques (en P2P), créez votre réseau:
- restreint en fonction de vos affinités
- ou étendu pour avoir une vue complète
adhérent
sympathisant
opposant
parti A
parti B
parti C
parti D

Les utilisateurs

Plusieurs utilisateurs à identifier (en tant qu'utilisateur, il ne s'agit que d'un identifiant sur le réseau, rien de nominatif). Rien ne doit permettre d'identifier une personne (malgré ce qui va suivre).

Les groupes d'utilisateurs (le réseau de chacun)

Cette notion de groupes d'utilisateurs (on pourra changer cette dénomination si on trouve un meilleur terme) est très importante car le dynamisme du réseau proviendra de ces groupes.

Un utilisateur qui se voudra ou prétendra être actif doit être capable de rattacher autour de lui plusieurs personnes aux idées similaires, et d'ouvrir ces questions à des réseaux un peu moins proches, mais avec des points communs.

Un exemple double : suivant les groupes, certains peuvent vouloir questionner sur le retour de la peine de mort et d'autres sur l'interdiction du pesticide tueur d'abeilles. Ces questions auront un impact auprès des personnes qui les émettent et qui les véhiculent. C'est en élargissant ces questions à d'autres groupes qu'il sera constaté si ces questions sont abandonnées ou gardent une certaine importance (même secondaire).

Cette notion de groupe est une particularité de Sondyn très importante : ce sont les groupes qui véhiculent les idées des sondages.

Il est possible et probable que ces groupes soient ultérieurement liés à des serveurs indépendants.

Une évolution possible ultérieure des groupes (hors cahier des charges)
Dans un second temps, nous pourrons imaginer devoir mettre en place, s'il y a une demande des utilisateurs, des hiérarchies de groupes : 1 = le ou les groupes auxquels on appartient, 2 = des groupes auxquels on n'appartient pas mais pour lesquels on accepte de recevoir des sondages, 9 = des groupes exclus dont on ne veut pas entendre parler.

Les données personnelles rattachées aux utilisateurs

Âge, sexe et CSP facultatifs
Au fur et à mesure de l'utilisation, des données pourront être collectées à la condition impératives qu'elles ne puissent pas être reliées à une personne physique. Parmi elles certaines sont facultatives et pas forcément conseillées, sauf à titre de fierté personnelle (ex : « Dans ma profession, on vote majoritairement pour tel candidat »). Ces données sont le sexe, l'âge ou la tranche d'âge, la catégorie socioprofessionnelle (CSP). Ces données étant facultatives, elles peuvent ne pas faire partie du cahier des charges et n'avoir été indiquées qu'à titre d'information.

La localisation géographique et les votes précédents
Cette demande sera facultative et on peut s'attendre qu'une grande majorité de personnes n'y répondent pas. En revanche, ces données sont primordiales pour pouvoir appliquer des correctifs : quand on interroge un échantillon sur ces intentions actuelles, et que l'on doit vérifier la représentativité de l'échantillon avec le vote précédent.

Exemple simplifié : dans une région donnée, on obtient sur Sondyn que 50 % des personnes interrogées vont voter pour le parti A et 30 % pour le parti B. Et ces mêmes sondés déclarent avoir voté précédemment aux dernières élections à 60 % pour A et à 25 % pour B. Or, dans les données publiques publiées après les élections, dans cette région A a obtenu 54 % et 28 % pour B (écart de 10%).
On corrigera donc les données du sondage en attribuant 45 % de prévisions à A et 33 % à B. (Il s'agit d'un exemple très simplifié.)

La connaissance des localisations et des votes précédents est très importante. Maintenant, par principe de ne pas trop collecter de données et de ne pas trop vouloir être intrusif, ce seront des données qui resteront facultatives (ou pas aussi précises que la ville, ne serait-ce que le département ou la région), en tablant sur une proportion plus grande de sondés et en augmentant notre marge d'erreur.

Ces données, lorsqu'elles sont collectées, devraient permettre également de mieux lutter contre les votes de bots.

Les données tierces d'autres réseaux sociaux
Par « données tierces » nous évoquons des indications volontairement données par certains utilisateurs (des données que nous ne devrions pas demander) et qui devront être prises en compte : ce sont les rattachements à d'autres réseaux sociaux qui seront indiqués comme une étiquette sur les comptes concernés. Ces données seront utiles pour les utilisateurs qui voudront faire connaître leur page Facebook ou Twitter mais pas utiles pour nous (aucune indication pour les sondages n'est à retirer).

S'il n'y a pas d'utilité en matière de sondage, ces rattachements pourront servir à authentifier des comptes et à valider des intentions de votes.

Nous insistons fortement sur le fait que les comptes ne doivent pas pouvoir être identifiés. Ces données tierces pourront donc par exemple être rattachées aux données de connexion du compte et de l'affichage des personnes qui proposent les sondages. Ce point reste à discuter.

Les réponses aux sondages

Les sondages ou questions posées par les utilisateurs
Il s'agit d'une grande différence avec les sondages des instituts traditionnels : ce ne sont pas les clients qui achètent des questions qui seront posées. Ceux qui voudraient être « clients » en quelque sorte doivent s'impliquer gratuitement dans le réseau et poser leur question auprès de leur entourage.

On doit donc s'attendre à une multiplicité de questions, qui peuvent être créées en doublons, par des personnes qui ne sont pas en contact.

Questions ouvertes et questions fermées
Les questions devraient être naturellement être plus souvent ouvertes (et donc avec de grandes listes de candidats fantaisistes qui pourraient apparaître) : « Pour qui envisagez-vous de voter à la prochaine élection ? »

Toutefois d'autres questions peuvent être fermées : « Parmi ces trois candidats, lequel représente le mieux selon vous les valeurs de [tel ou tel parti] ? »

Les remontées des questions
C'est l'intérêt du projet : pouvoir faire une synthèse des différents avis pour des questions identiques, et renvoyer le résultat global aux utilisateurs qui ont posé cette question à leur réseau.

On retourne, pour chaque initiateur d'une question posée, les résultats de son groupe de contact ainsi que les résultats plus globaux.

L'application des correctifs lors de la remontée des résultats reste à définir.

La durée de validité des réponses
Les réponses doivent avoir une durée de validité (à la différence des instituts traditionnels qui posent des questions ponctuelles sur 1, 2 ou 3 jours par exemple). Par défaut, on proposera un mois, tout en précisant que la personne peut venir et modifier son choix à tout moment si elle change d'avis.

Cette durée de validité est importante afin de ne pas fausser les résultats avec une personne qui a donné un avis il y a trop longtemps.

On aura donc des questions sans date de fin (ou jusqu'à une lointaine élection), et des réponses attachées aux personnes qui y répondent, sur une période définie.

Nous aurons des résultats sur une période glissante.

Les réponses données, une fois la validité dépassée, ne sont pas effacées mais restent sur le réseau.

Les mises en relation de personnes

Rappel très important : nous ne voulons pas et nous ne devons pas avoir accès aux réponses des gens sur des sondages. Pour une proposition de mise en relation de personnes, on n'enverra pas de message du genre « Vous avez voté comme telle personne, voulez-vous être mis en relation ? », c'est, d'un point de vue éthique, hors de question.

En revanche, des suggestions de rapprochement pourront être faites en fonction de sondages similaires publiés.

D'autres suggestions pourront être faites par rapport à un nombre d'« amis » communs.

Les enregistrements de données

Les données sont enregistrées sur le premier serveur sur lequel la personne est connectée. Une fois l'enregistrement effectué, une réplication de ses données est faite sur les serveurs en liaison avec le premier serveur, tout comme d'autres opérations de mise à jour.

Ce qu'il n'est pas prévu d'avoir

Par exemple : pas de forums ou des discours de présentation des idées. Leur place est sur les réseaux sociaux existants (Twitter, Facebook, Mastodon...), là où un lien pourra être mis en direction de Sondyn (s'il n'est pas censuré par des gros genre Facebook ; ce qui apparaît malheureusement possible). Nous n'affichons que des sondages.

De même, il n'est pas prévu d'ouverture vers Fediverse ou ActivityPub.

Question de méthodologie des sondages, il pourrait exister des questions sur la certitude de voter (du style « Notez votre certitude de voter pour x ») , mais ce n'est pas à l'ordre du jour.

ÉTAPE INTERMÉDIAIRE

Le programme sera axé sur du pair à pair entre des serveurs individuels. En attendant que le programme soit pleinement fonctionnel, un site temporaire centralisé classique sera créé en parallèle sur un serveur unique style PHP-MariaDB (je m'en charge moi-même, Lionel Aubert, indépendamment du présent cahier des charges).

Au fur et à mesure de l'avancement du programme, des pans du système temporaire intermédiaire seront basculés vers l'environnement final.

Il s'agit d'une possibilité qui reste à discuter (le moment du basculement de pans du programme).

Délais

Aucune précision au stade actuel. Ça dépendra de la vitesse d'avancement du projet, tout en considérant qu'avec un système centralisé dans un futur proche, il y aurait moins d'urgence.

Toutefois, si ce programme pouvait être prêt avant les élections départementales et régionales de 2021, ce serait une bonne chose.

RETOUR SUR LES DONNÉES COLLECTÉES ET TRAITÉES

(Il s'agit de propositions qui pourront être modifiées pour d'autres choix techniques)

Il s'agit de monter à terme un réseau pair-à-pair décentralisé hybride, avec des serveurs intermédiaires de type super-pair tandis que les utilisateurs sont des pairs simples.

Le réseau

À ce stade, pas d'indication sur le moyen de se connecter au réseau (à étudier et à définir).

- Identifiant de la personne qui participe au programme de sondages (= « participant »).
- Identifiant d'un réseau sur lequel on se connecte et qui conserver les données (= « serveur »).

Remarque :
Le terme de « serveur » est ici une simplification de langage, car tous les nœuds sont en réalité des clients-serveurs. Simplement certains ne seront en service que temporairement et avec des capacités de stockage et de travail limitées (les participants, avec leur simple application sur leur téléphone pour l'essentiel), tandis que d'autres décideront de faire tourner des machines plus conséquentes (un ordinateur fixe par exemple), quelques fois en continu malgré un impact écologique défavorable. Ces machines seront des « super-pairs ». Leur fonction essentielle étant de se substituer à un serveur central, de là, par simplification de langage, on pourra les considérer comme des serveurs intermédiaires. Le réseau sera de type pair-à-pair hybride.

Paire de valeurs participant-réseau (tables de hachage à définir).

- Identifiant des données du participant (qui devraient être en partie disjointes des données du participant).

Les données du participant doivent être dupliquées sur plusieurs serveurs et mises à jour quand les serveurs sont connectés entre eux

mise à jour

Les données d'un participant doivent être incrémentées avec une date. Lors d'une vérification de mise à jour entre serveurs, on vérifie pour chaque participant le nombre de champs enregistrés et la date de la dernière modification. On pourrait aussi vérifier l'ensemble des données avec un SHA-256.
Si les enregistrements de part et d'autre ne correspondent pas, on les met à jour, en cascade (avec éventuellement des duplications ou des effacements de données).

Duplication

Le principe étant de disposer des données même lorsque des utilisateurs ne sont pas connectés, il convient de les dupliquer en nombre suffisant sur plusieurs postes super-pairs. Ce nombre est à définir. Pour un exemple de départ, il pourrait y avoir 2 valeurs, l'une de jour, l'autre de nuit, respectivement de 15 à 20 et de 5 à 10. (Nous gardons à l'esprit que les ordinateurs qui servent de super-pair peuvent être éteints la nuit.)

Effacement

Nous précisions que la totalité des valeurs de sondages à des dates passées seront toujours conservées. Lorsque nous parlons d'« effacement », il s'agit de supprimer des réplications de données sur le réseau trop importantes.

Lors d'une mise à jour de données, si nous nous apercevons d'une présence en sur-nombre, nous pourrions envisager un effacement des données sur-numéraires, par exemple chez des pairs qui ne constituent pas un bout de chaîne.

Ces effacements seront traités en seconde partie de la réalisation du programme.

Les données des utilisateurs

Les données des utilisateurs sont distinctes de celles des résultats des sondages.

Ce sont les données qui permettent à l'utilisateur de se connecter et d'informer le réseau de sa présence. Ça sera son adresse courriel, son pseudonyme ou son nom, son mot de passe de connexion et éventuellement d'autres informations qu'il aura décidées librement comme des liens vers ses comptes sur des réseaux sociaux). Il faudra veiller à une protection optimale de ces données.

Toujours pour des raisons de sécurité, nous pourrions imaginer une réplication de ces données à condition que les postes sur lesquels elles seront répliquées ne contiennent pas les données des sondages de la personne. Ces postes devront correspondre à une sorte d'hôtes de confiance (ce qui restreint le choix des super-pairs).

Une autre possibilité serait d'imaginer que ces données des utilisateurs soient volontairement coupées en plusieurs morceaux (2 ou 3 par exemple) répartis sur plusieurs postes distincts, et que seule la connexion de l'utilisateur permettrait de joindre. Il ne s'agit, à ce stade, que d'une proposition (et il peut y en avoir plusieurs autres).

L’étude de ces solutions techniques sécurisées occupe une part importante dans les solutions demandées par ce cahier des charges.

Les données des utilisateurs sont des données modifiables (changement de pseudonyme, de mot de passe, de liens vers les réseaux sociaux…), tout comme elles pourront être supprimées lors de la fermeture d’un compte demandée par l’utilisateur.

Les données des réponses aux sondages

Les données relatives à des sondages s’incrémenteront au fur et à mesure. Les données correspondent aux identifications des différents sondage, à leurs réponses et la date. S’ajouteront les données permettant de donner un coefficient aux sondages (votes lors des élections précédentes). Aucune de ces données ne sera effacée ni modifiée, pour conserver un historique et affiner les résultats globaux.

Ces données ne doivent en aucune façon être liées ou pouvoir être liées au participant aux sondages (la seule manière d’y accéder est depuis le compte utilisateur, qui rajoute le cas échéant des données supplémentaires). Ainsi, que le compte d’origine du sondage existe ou ait été supprimé est sans incidence. Ce qui importe, c’est la préservation de ces données pour pouvoir être traitées.

Les questions posées (sondages)

Ce sont des données indépendantes. Le nom du ou des créateurs des questions peut être mentionné (sans pouvoir être lié à ses données personnelles).

Une table de hachage sera utilisée pour retrouver les sondages.

Les calculs effectués sur les réponses aux sondages

Les différents calculs seront à recopier et à adapter à partir du site temporaire en cours de création.

GRAPHISME

Pas très important dans un premier temps.

BUDGET

Le budget sera collecté au fur et à mesure par la recherche et la venue de nouveaux investisseurs. Une première enveloppe entre 20 et40 000 euros est espérée.

Les embauches ne sont lancées que lorsque 2 mois de paiement sont en réserve ou sur le point de l’être.

DÉLAIS

Le souhait serait un programme opérationnel quelques mois avant les élections de mars 2021. Toutefois, grâce au programme temporaire, un dépassement ne devrait pas avoir de grande incidence.

HÉBERGEMENT

À voir au fur et à mesure de l'avancement.

LUTTE CONTRE LES BOTS

C'est problématique à deux niveaux :

- surcharge d'utilisateurs inscrits (saturation du système)
- données faussées (intentions de vote)

On pourrait imaginer

ÉCOLOGIE

Un réseau de plus en P2P, et surtout s'il incite des gens à laisser leur propre ordinateur tourner plus souvent n'est pas bon pour les questions environnementales (empreinte carbone, etc.). Toutefois, si ce programme permet, par une réorientation citoyenne des personnes sondées, une meilleure prise de conscience de ces questions ou même de questions parallèles telles que la mortalité des abeilles, la disparition massive de forêts équatoriales (Mercosur), etc, alors l'impact négatif du réseau aura permis, dans d'autres domaines, des avancées.

COOKIES, AUTRES COLLECTES D'INFORMATION

Nous serons très attentifs à nous limiter au strict minimum.

Nous ne collecterons surtout pas des numéros de téléphone.

AUTRES INFORMATIONS

Ce cahier des charges n'est qu'une ébauche au 22 août 2020 et sera amélioré dans les jours ou semaines à venir. Merci pour votre lecture.

Réalisé par Lionel Aubert, pour le projet Sondyn.