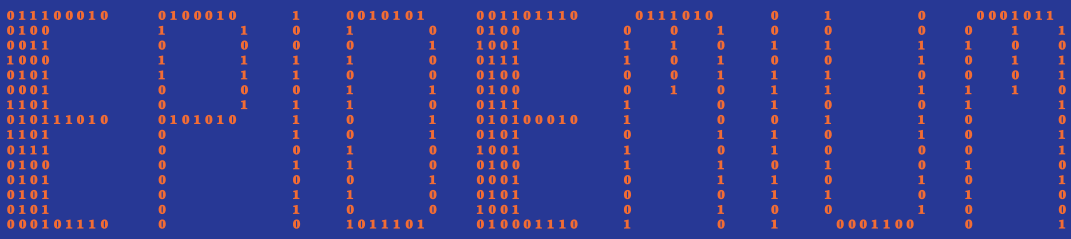
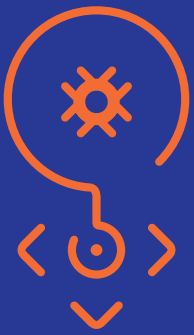




LIVRE BLANC

OPEN SCIENCE | ÉPIDÉMIOLOGIE DU CANCER | BIG DATA | COMMUNAUTÉ

Ouvrage coordonné par l'équipe Epidemium :

- Mehdi Benchoufi
- Olivier de Fresnoye
- Karine Lévy-Heidmann
- Ermete Mariani
- Ozanne Tauvel-Mocquet

// **Crédits :**
- Logo et pictogrammes Epidemium : Marie-Sarah Adenis et Lucile Picon
- Dessins : concept Ermete Mariani et design Barbara Govin
- Conception graphique et mise en page :  www.ediconcept.fr
- Imprimé en France - Janvier 2017

// **Édition imprimée :** ISBN 979-10-97214-00-5
// **Édition numérique :** ISBN 979-10-97214-01-2

Dédicace

Nous dédions ce livre à tous les patients qui ont été à l'origine du projet et à toutes celles et ceux qui voudront s'engager et contribuer à faire grandir Epidemium.

Sommaire

Préface

Pr Cédric Villani

Parfois le progrès réside dans la simple amélioration, où l'on apprend à faire les choses mieux, plus vite, plus efficacement. Et parfois il passe par un changement plus radical où les habitudes mêmes sont bousculées ; alors émergent non seulement de nouvelles techniques, mais aussi de nouvelles organisations du travail.

Aujourd'hui l'une des techniques émergentes qui agitent le plus le monde est l'analyse de grands jeux de données. Pas si facile de dire exactement de quoi il s'agit ; d'ailleurs, même le nom de la discipline n'est pas clair : big data, *data mining*, mégadonnées ? Pas facile non plus de définir son périmètre, puisqu'il s'agit d'un mélange de statistiques, analyse, géométrie, probabilités, optimisation, ... Mais l'enjeu est clair : on attend tout et encore plus de ces techniques, qui ont déjà révolutionné la recherche d'information, la traduction automatique, l'intelligence artificielle, et les modèles économiques de bien des entreprises, dont les géants emblématiques de l'Internet. On attend tellement de l'analyse de grandes données que le blason des statisticiens a été couvert d'or fin, et que ce métier est devenu l'un des plus demandés au monde - en 2015, métier d'avenir numéro 1 selon la compagnie américaine Careercast.

Il était donc évident que l'analyse de grandes données allait s'attaquer, un jour ou l'autre, à l'un des fléaux qui sévissent encore le plus dramatiquement dans le monde, en fait l'un des plus grands sujets d'inquiétude dans les pays développés : le cancer. Quelle famille, dans notre pays, n'a pas été touchée par cette maladie ? Fléau d'autant plus redoutable qu'il est multiforme, varié, que ses causes et facteurs de risque sont extraordinairement multiples.

Et c'est justement pour cela que l'on attend tellement de l'alliance entre grandes données et cancérologie : il y a tant de statistiques mais elles sont si difficiles à interpréter, si variables, que l'on se dit qu'il faudra forcément utiliser des méthodes nouvelles pour en venir à bout et découvrir des choses intéressantes, de nouveaux facteurs que les médecins pourront mettre en œuvre.

Mais dans le projet Epidemium, il y avait aussi l'idée que cette nouvelle technique devait s'accompagner d'une nouvelle façon de travailler, publique et ouverte, à l'image des méthodes qui ont fait le succès du logiciel libre dans les années 90, et celui des FabLabs dans les années 2000. Une organisation où les notions de plateforme, Wiki, échange de données, partage du travail entre organisations, coopérations bénévoles, synergies, prendraient tout leur sens ; où l'on partagerait un cadre fait d'outils, de facilités, de compétences ; et où l'on laisserait la compétition faire le tri entre les idées.

Emblématique de ce projet était le rapprochement entre La Paillasse, organisation

militante pour une recherche ouverte, et Roche, poids lourd de l'industrie pharmaceutique ; symbole de ce que les grandes institutions de recherche ont senti le potentiel qu'il y avait dans la recherche médicale ouverte.

Les fondamentaux d'Epidemium étaient posés, il restait encore tant d'obstacles à franchir ! Recenser les bases de données, définir un concours, rassembler largement les énergies, recruter un jury ; mais aussi définir les principes qui encadreraient ce concours.

Soucieux de suivre les bonnes pratiques, les organisateurs recrutèrent un Comité d'éthique indépendant, à qui il reviendrait de définir des garde-fous dans l'usage des jeux de données et que l'on pourrait consulter pour des questions délicates, des dilemmes à résoudre. Car la mise en relation de bases de données comporte aussi bien des promesses d'efficacité accrue que des craintes d'intrusion inacceptable. C'est à ce comité que j'ai eu le plaisir de participer : tâche légère mais importante, qui a aussi été source de réflexion quand il s'est agi d'évaluer les projets en compétition.

Une autre bonne pratique était l'implication d'une « association de malades » car les patients ont leur mot à dire, bien sûr, ne sont-ils pas les premiers concernés ? Il était donc légitime de leur laisser une part importante dans la gouvernance.

Et surtout, il y avait une grande volonté de synergie entre d'une part les experts de l'abstraction (mathématiciens, statisticiens, informaticiens) et d'autre part les experts du corps humain (cancérologues, médecins, chirurgiens, ...) une synergie à mettre en place entre personnes, pour refléter la synergie entre disciplines ; un effort dont les résultats ne pourront pleinement s'apprécier qu'à travers la coopération de long terme, et auquel Epidemium a souhaité donner un coup de fouet.

Dès le top départ, que de travail a été accompli ! C'était fascinant d'assister, de loin, à l'activité dont faisaient preuve les équipes en compétition, dans un grand chaos productif.

Et le moment venu, il a fallu écouter, départager, remettre des prix... C'était la fin d'une étape mais il était clair pour tous que c'était surtout le début d'une aventure de longue haleine, et qu'il allait falloir continuer à capitaliser sur ces acquis pour participer à l'émergence de la médecine du futur. ■

Avant-propos

Isabelle Vitali (Roche) & Thomas Landrain (La Paillasse)

Epidemium : innover les pratiques de la recherche dans l'épidémiologie du cancer

Innover, ce n'est pas seulement développer de nouvelles technologies ou des objets connectés. C'est aussi expérimenter de nouvelles dynamiques et méthodologies de travail pour aider au développement du potentiel de l'intelligence collective, tout en dépassant les barrières qui cloisonnent les savoirs et les compétences. De plus, nous pensons sincèrement que l'innovation doit avoir des retombées positives et tangibles sur les conditions de vie des personnes et sur leur santé.

C'est sur cette base commune que la rencontre entre le laboratoire de recherche ouverte La Paillasse et l'entreprise pharmaceutique Roche a pu donner vie, grâce notamment à l'entremise de Gilles Babinet, à un projet très ambitieux : Epidemium.

Né en avril 2015, le programme Epidemium représente l'effort conjoint de Roche et de La Paillasse d'expérimenter une manière innovante de faire de la recherche scientifique et médicale, sur un problème majeur de santé publique comme le cancer, en utilisant une source de connaissances en plein essor bien qu'encore sous-exploitée : le big data. Il nous tenait aussi à cœur de prouver que, dans le périmètre de la science, la rencontre d'acteurs de nature différente et animés par les mêmes principes est source d'innovation.

Le choix de lancer un projet novateur, tant sur la forme que sur le fond, en épidémiologie du cancer, nous a semblé logique pour deux raisons principales :

• malgré les avancées scientifiques ré-

centes, le cancer est responsable de 8,2 millions de décès en 2012, dont 148 000 en France[1], avec la prévision terrifiante de 60% de décès en plus chez les femmes d'ici 2030 dans le monde ;
- le big data offre une source précieuse de connaissances pour la recherche médicale, mais ses potentialités demeurent largement sous-exploitées dans le secteur de la santé.

Pour qu'un tel projet soit un succès, nous avons immédiatement compris qu'il fallait réunir une équipe insolite, composée de personnes aux compétences très différentes, ayant des parcours et venant d'horizons parfois éloignés, mais qui partagent et défendent tous les mêmes valeurs fondamentales : l'ouverture du savoir, la collaboration, la transdisciplinarité et l'indépendance.

> *Dans le monde de la techno-médecine 2.0, Big Pharma et biohackers s'allient contre le crabe ! »*
>
> ***Dominique Nora***
> *(L'Obs, 05/11/2015)*

Dès le début, nous avons eu l'intuition, qui fut confirmée par la suite, que ce projet nous porterait vers des territoires encore inexplorés de la science et de l'éthique. Il nous fallait donc veiller à accompagner la communauté dans ce parcours et c'est dans cette optique que nous avons créé le Comité d'éthique indépendant et le Comité scientifique. Ces deux comités ont fait face à des problématiques inédites qui dépassent le périmètre d'un projet de recherche traditionnel. C'est grâce à leur engagement constant que nous avons réussi à trouver le juste équilibre entre les nécessités de la recherche scientifique et la sauvegarde des droits ainsi que du bien-être des patients en particulier, et de tous les citoyens en général, qui sont toujours au centre de notre engagement.

La communauté si hétérogène et étendue d'Epidemium, plus de 1 000 personnes ayant participé à différents titres à nos évènements pendant un an, n'aurait pu être productive sans le suivi d'une équipe de coordination, qui a su mobiliser un vaste réseau de parties prenantes, partenaires et membres de la communauté.

Dès ses débuts, Epidemium a suscité l'intérêt de plusieurs acteurs du milieu médical, de la recherche scientifique et du big data, qui se sont reconnus dans nos valeurs et qui partageaient notre ambition. Cet intérêt s'est vite transformé en partenariats très importants pour la réussite du projet et qui ont permis de mettre à disposition de la communauté compétences, outils et ressources.

Après une première année de programme, le temps est donc venu de faire un bilan objectif et de partager avec vous, en toute humilité, notre expérience et les enseignements que nous en tirons pour continuer la dynamique lancée avec Epidemium.

Epidemium nous a donné l'opportunité, chez Roche et La Paillasse, de mettre en synergie nos compétences, de découvrir des complémentarités inattendues entre nos deux organisations, mais surtout de nous laisser surprendre par les résultats de l'intelligence collective exprimée par

la communauté ouverte que nous avons supportée et accompagnée.

En tant qu'acteur de l'innovation en santé, l'ambition de Roche est, avec Epidemium, de réinventer l'épidémiologie du cancer pour en faire un véritable outil d'aide à la décision thérapeutique et de porter l'innovation au plus proche des patients. En regardant, après plus d'un an, les résultats atteints, nous pouvons mesurer l'immense pas en avant fait dans cette direction. De plus, Epidemium a été pour Roche une occasion unique d'expérimenter de nouvelles méthodes et de nouveaux instruments pour stimuler l'innovation dans le cadre de la science ouverte, jusqu'à ouvrir certaines de ses propres données à la communauté.

Lieu emblématique d'innovation et de transdisciplinarité, La Paillasse défend l'idée de faire émerger une nouvelle manière de faire de la recherche, grâce à la mise à disposition d'un environnement de travail ouvert et collaboratif. La Paillasse est le lieu où s'ancre l'histoire d'Epidemium, lieu qui a permis à la communauté de grandir en s'identifiant à un espace réel, élément fondamental pour garder le juste équilibre entre les échanges virtuels et en face à face. Epidemium a aussi été l'occasion pour La Paillasse d'avancer dans le perfectionnement de ses outils et méthodes de travail, ainsi que de montrer le potentiel de l'intelligence collective appliquée à la recherche médicale.

Nous voulons également profiter de l'occasion offerte par ce *Livre blanc* pour partager avec vous notre vision de ce qui pourrait être amélioré. En effet, ce que nous avons initié ne s'arrêtera pas là…

Le premier de ces points, et c'est ce que nous nous efforcerons de faire pour les années à venir, est d'impliquer davantage la communauté médicale. Notre objectif de départ était et reste d'obtenir des résultats et des outils concrets qui pourront véritablement être portés par les équipes médicales dans leur travail quotidien. Or, cela ne peut se faire que si les médecins et l'ensemble du monde médical s'engagent de manière constante tout au long du programme. Certes, travailler sur les données et développer des algorithmes de *machine learning* est essentiel pour faire progresser la santé aujourd'hui, et c'est ce que nous avons voulu mettre au cœur d'Epidemium. Mais cela ne peut se faire efficacement sans les apports d'experts de la santé et de cliniciens.

> *« Nous désirons plus que tout que le programme continue à se développer. »*

De plus, nous souhaitons améliorer l'intelligibilité du programme et de ses enjeux. Nous sommes conscients, après un an de programme, que le sujet de la santé alimenté par le big data est complexe et qu'il existe probablement d'autres façons de le traiter. Nous nous efforcerons toujours d'adopter une approche collaborative et ouverte dans l'appréhension du programme ainsi que dans la manière dont nous le construisons. Le moyen que nous avons choisi pour aboutir à nos objectifs aurait pu être différent ; il doit certainement être amélioré.

C'est également la raison pour laquelle il était important pour nous de créer ce *Livre*

blanc, afin de prendre du recul sur le travail accompli, de garder une certaine objectivité sur ce qui a été réalisé et de se donner des perspectives pour l'avenir puisque nous désirons plus que tout que le programme continue à se développer.

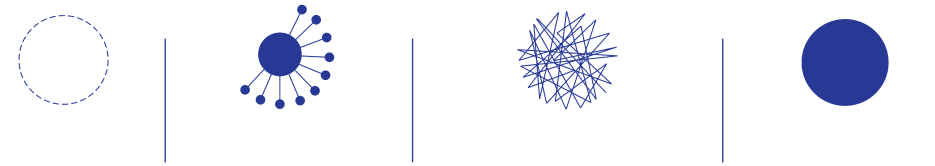
Soyons reconnaissants : beaucoup de ce qui a été réalisé nous a épatés. Nous avons été bluffés par la qualité des rendus, par la capacité d'une communauté à se mobiliser sur un sujet comme le cancer, par la qualité des experts qui sont restés engagés tout au long du programme et enfin par la bienveillance générale qui a entouré le programme, que ce soit de la part des participants, des partenaires, des comités ou encore du grand public. ■

1. Unicancer < www.unicancer.fr >, dernière consultation le 30 novembre 2016.

Introduction

Équipe Epidemium

LES VALEURS DU PROGRAMME



Epidemium est né d'une volonté : travailler en cancérologie sur la base des données ouvertes en adoptant une approche communautaire. Celle-ci reposait alors sur deux intuitions. La première, qu'il est possible d'obtenir des résultats pertinents en décloisonnant la recherche médicale ou, du moins, en la rendant accessible à un plus grand nombre d'acteurs, y compris non scientifiques. La deuxième, qu'il est possible de fonder une recherche sur l'*open* big data, étant persuadés que ces dernières offrent de nouvelles perspectives dans la compréhension de notre environnement et de nous-mêmes : mieux comprendre, mieux prévenir, mieux prédire. Ainsi, Epidemium a dû mettre en place une méthodologie visant à concilier recherche scientifique, communauté et données ouvertes, ces trois éléments représentant en quelque sorte l'ADN du programme.

Partant de ces intuitions et de cette volonté initiales, Epidemium s'est naturellement pensé comme un programme d'*open science* et s'est construit autour de quatre valeurs fondamentales : l'ouverture, la collaboration, la transdisciplinarité et l'indépendance. Concrètement, ces dernières se sont incarnées à travers les modalités d'un programme accessible à tous, qui prône et facilite le partage des méthodes, des connaissances générées, le travail collaboratif et l'échange de compétences ainsi que la transdisciplinarité des profils. Enfin, le programme est autonome vis-à-vis de ses initiateurs Roche et La Paillasse, et s'est pour cela entouré d'un Comité d'éthique indépendant. Toutefois, qu'est-ce véritablement que l'*open science* et en quoi le programme Epidemium est-il innovant ?

Traditionnellement, les recherches scientifiques se réalisent au sein de structures dé-

diées, cloisonnées, dans lesquelles la production de la connaissance est encadrée et sa diffusion limitée. Cependant, depuis quelques années, de nouvelles structures de production de la connaissance scientifique émergent, caractérisées par deux éléments majeurs : d'une part, le nombre de participants aux projets, d'autre part, l'ouverture des problèmes et des résultats à toutes les étapes de la production. Les barrières tombent, aussi bien dans l'intégration de nouveaux acteurs aux projets qu'au niveau de la propriété intellectuelle. Chaque « citoyen de la science » peut intervenir et participer au développement et à la résolution des problèmes en apportant ses connaissances et ses compétences. Nous retrouvons chez les participants aux projets d'*open science* des valeurs communes et des idéaux qui rassemblent au-delà même de la thématique du projet, dans un souci de partage des découvertes. C'est l'avènement d'une nouvelle conception épistémologique de la recherche scientifique. La dynamique de l'*open science* favorise les interactions entre les différents acteurs du projet, ce qui améliore d'autant plus la capacité du collectif à générer des solutions puissantes et originales. Ainsi, pour Epidemium, l'accès à des compétences et des connaissances variées offre statistiquement une plus grande diversité dans l'approche des problématiques autour du cancer ainsi que de nouveaux points de vue, qui sont autant de directions potentielles à explorer. De plus, l'ouverture des résultats au fur et à mesure du projet permet à tout contributeur d'accéder aux dernières avancées, ce qui le rend capable à tout moment de rejoindre le projet, de proposer des alternatives ou d'améliorer celles proposées, rendant à nouveau plus fortes les perspectives pour la recherche.

> *Epidemium a tout pour prouver que la science peut se faire en dehors des cadres académiques classiques en misant sur l'open source et la transdisciplinarité d'équipes auto-constituées. »*
>
> ***Hugo Jalinière***
> *(Sciences et Avenir, 30/05/2015)*

La volonté de concevoir Epidemium comme un programme d'*open science* ne découle pas seulement de l'ambition de ses deux initiateurs. L'idée de travailler sur le cancer de façon innovante, à la fois par la forme du programme et par la méthode proposée, alliant ouverture et utilisation du big data, est née de plusieurs constats sur la maladie et le contexte actuel. La thématique du cancer est contemporaine, porteuse de sens et fédératrice. À l'échelle mondiale, en 2012, 8,2 millions de personnes sont décédées des suites d'un cancer et cette incidence est amenée à augmenter de 70% au cours de la prochaine décennie[1]. Chacun d'entre nous est ainsi touché par cette maladie au cours de sa vie, de près ou de loin. Le cancer est donc un enjeu de société majeur, qui entraîne des réactions émotionnelles fortes et a des effets palpables. De nombreuses communautés de patients, de proches ou d'acteurs de la santé, se sont en outre déjà construites spontanément dans le cadre de la recherche contre le cancer et pour défendre différents intérêts. Epidemium offre alors la possibilité de fédérer tous ceux qui le souhaitent dans un grand mouvement de recherche commun.

Epidemium propose à des acteurs habituellement peu sollicités à ce niveau-là des moyens pour se regrouper, et permet leur *empowerment* grâce à une facilité technique : l'accessibilité des données et la démocratisation des outils de traitement. Des données ouvertes et hétérogènes sont disponibles, notamment sur les sites intergouvernementaux, et en quantité suffisamment importante pour en induire du sens, en dégager des pistes de recherche. Or le big data impacte naturellement l'épidémiologie du cancer et c'est peut-être l'un des domaines où il est susceptible d'être le plus porteur de sens : il propose des données concernant toutes les dimensions des sociétés, du quotidien des individus et de l'environnement. Le big data, par sa complexité et les possibilités qu'il offre, demande un niveau de transdisciplinarité important de la part des acteurs qui vont le traiter, l'étudier et en tirer des conclusions. C'est pour cela qu'Epidemium est ouvert à tous ceux qui souhaitent mettre leurs compétences, quelles qu'elles soient, à disposition de ce sujet.

> *« Avec Epidemium, on a réussi à prouver qu'il était possible, en France, de réunir des personnes brillantes et motivées pour produire de la science de qualité, et de réunir des experts dans le Comité scientifique et le Comité d'éthique indépendant, pour guider et évaluer les projets. »*
>
> ***Dr Charles Ferté***
> *Membre du Comité d'éthique indépendant*

Forts de l'expérience du Challenge4Cancer et convaincus de l'intérêt d'un programme tel qu'Epidemium, nous avons souhaité, afin de clore cette première édition et d'imaginer la prochaine, rédiger ce *Livre blanc*. Fidèle, dans sa conception, aux valeurs que nous avons défendues, celui-ci fait écho à la pluralité des points de vue et des disciplines engagées, en recueillant les recommandations des acteurs qui ont pris part au programme : membres de la communauté, membres des comités d'éthique et scientifique, contributeurs et partenaires. Ainsi, ce livre se veut être un plaidoyer à la fois pour l'*open science* et pour la méthodologie collaborative.

C'est un objet composite que nous proposons, regroupant des articles, parfois co-signés, des retours d'expérience et des fiches d'approfondissement, tous construits autour des thématiques fondatrices du programme, à savoir la santé, l'*open* et les données. Libre au lecteur de le lire de façon linéaire ou de choisir d'explorer les sujets qui lui plaisent ou qui lui parlent. De cette multiplicité, trois parties ont émergé autour desquelles ce *Livre blanc* est structuré : *La Communauté agile et ouverte* qui présente la méthodologie employée et fait remonter le vécu de la communauté ; *L'Innovation scientifique et médicale* qui ouvre une discussion sur la rencontre de la *data science* et de la médecine, sans oublier le bénéfice patient qui se pose toujours en ligne d'horizon ; et enfin *Le Cadre juridique et ouvert* qui revient sur les problématiques à la fois légales et éthiques qu'a pu poser et rencontrer Epidemium dans sa mise en place et réalisation. ■

1. Centre international de Recherche sur le Cancer (CIRC) - Organisation mondiale de la Santé (OMS), Communiqué de presse n°223, <www.iarc.fr>, dernière consultation le 30 novembre 2016.

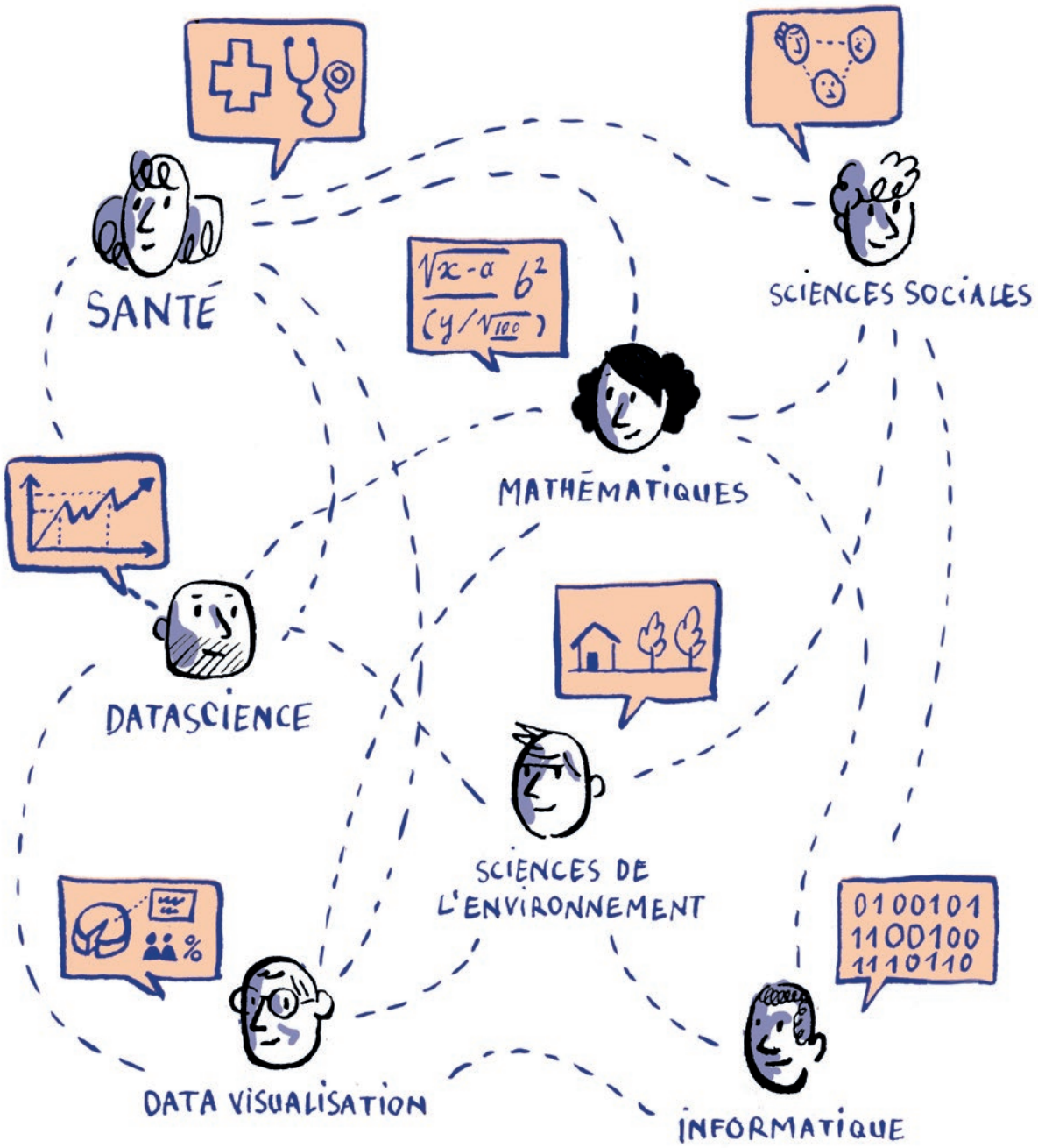
#1

UNE COMMUNAUTÉ AGILE ET OUVERTE

C'est grâce à la création et au maintien d'une communauté bénévole active qu'Epidemium a pu s'attaquer à un défi aussi ambitieux que celui de réunir les termes santé, data et open pour faire avancer la recherche contre le cancer. Quelles ont été les bonnes et les mauvaises pratiques dans l'organisation de la vie de cette communauté, tant au niveau de la compréhension de son fonctionnement que de son animation pendant les six mois de Challenge ? Quels enseignements en tirent les acteurs principaux du programme pour poursuivre avec un nouvel élan ce défi ?

// AUTEURS

Équipe Epidemium / Djalel Benbouzid / Léo Blondel / Marc Santolini Stéphanie de Haldat / Thomas Landrain


SANTÉ
SCIENCES SOCIALES
MATHÉMATIQUES
DATASCIENCE
SCIENCES DE L'ENVIRONNEMENT
DATA VISUALISATION
INFORMATIQUE
0100101
1100100
1110110

Allier cancer et big data grâce à une méthodologie agile

EPIDEMIUM CHALLENGE4CANCER (C4C) RAMP

COOPÉTITION OPEN SCIENCE

Comment travailler en cancérologie sur la base de l'open big data en adoptant une approche communautaire et ouverte ? Dans cet article, nous proposons de présenter et d'expliquer la méthodologie qui a sous-tendu la conception du programme Epidemium et la mise en place du Challenge4Cancer. Par cette méthodologie, nous avons notamment cherché à répondre aux problématiques posées par l'association des termes santé, data et open tout en tentant de favoriser la formation, le maintien et le dynamisme de la communauté, tout au long des six mois du Challenge.

// **AUTEURS**

Mehdi Benchoufi / Olivier de Fresnoye / Karine Lévy-Heidmann
Ermete Mariani / Ozanne Tauvel-Mocquet

#1

UNE COMMUNAUTÉ AGILE ET OUVERTE

Introduction

Animer une communauté aussi étendue et hétérogène que celle d'Epidemium représente un défi constant pour l'ensemble de l'équipe coordinatrice. Dès le début, il nous a paru évident que nous devions être suffisamment agiles pour stimuler la créativité d'une communauté en constante évolution, composée de personnes aux parcours et horizons très éloignés ainsi que de parties prenantes et de partenaires qui n'avaient pas l'habitude de collaborer. Notre mission était d'autant plus complexe que l'objectif du programme était de faire se rapprocher l'univers de la santé et celui du big data, dans un cadre ouvert de partage des connaissances, avec une mission précise : appréhender l'épidémiologie du cancer autrement.

La première étape de ce parcours a été d'organiser le Challenge4Cancer (C4C), à la façon d'une véritable compétition d'une durée de six mois et avec une remise de prix finale, mais où primerait plutôt la collaboration et le partage. En bref et pour utiliser un néologisme du milieu *open*, il s'agissait d'un défi communautaire coopétitif[1]. Pour cela, nous avons préféré structurer le C4C autour de quatre grandes thématiques, sans fixer d'objectif précis à atteindre. Au sein de ce Challenge, ouvert à tous les possibles quant à la nature et à l'objet des projets, les équipes avaient donc pour seules contraintes de respecter le cadre éthique et méthodologique établi par les comités ainsi que le règlement du Challenge.

Nous avons également voulu le rendre le moins contraignant possible et intellectuellement très stimulant, avec le souci de maintenir un engagement suffisamment fort de la part des participants tout au long des six mois, et principalement sur leur temps libre. Notre ambition était que cet engagement se concrétise en des projets documentés, qui devaient être évalués à terme par un jury constitué des comités d'éthique et scientifique. Or, il nous a fallu composer avec un engagement plus ou moins fluctuant, selon la motivation et les disponibilités des participants. De plus, en mobilisant une communauté large et hétérogène, il était aussi nécessaire de réfléchir à la façon de faire avancer tous les projets à une vitesse à peu près équivalente afin de maintenir une forme d'émulation et de

> *Epidemium , un programme de recherche participatif unique au monde. »*
>
> ***Jean-Bernard Gallois***
> *(01Net, 18/11/2015)*

Allier cancer et big data grâce à une méthodologie agile

pouvoir répondre aux besoins de chacun plus efficacement. À cela se sont ajoutés des points plus techniques, liés aux contraintes géographiques et temporelles. Il a fallu penser le Challenge de façon à ne pas rendre l'éloignement physique discriminant, afin d'accueillir les participants qui venaient non seulement de toute la France, mais aussi de l'étranger.

Enfin se posèrent deux enjeux cruciaux pour la réussite du Challenge, à savoir l'hétérogénéité des profils et, par extension, celle des compétences de chaque participant. Comment rendre le C4C accessible aux plus novices tout en restant attractif pour les plus expérimentés ? Comment éviter le risque que les premiers soient découragés et les seconds lassés ? Les profils médicaux connaissent les problématiques actuelles liées au cancer, les besoins des patients et l'état de la recherche ; et les profils liés à la science des données détiennent le savoir-faire lié au big data ainsi qu'une culture propre pour traiter les données, les interpréter et innover à partir elles. Comment alors favoriser les rencontres de tous ces profils différents mais complémentaires, pour que les projets soient en capacité de proposer des solutions globales et viables ?

Nous avons tenu compte de ces questions clefs pour organiser le programme et pour penser le fonctionnement du Challenge4Cancer ainsi que les outils mis à la disposition des participants.

Une communauté qui répond aux besoins du programme et de ses objectifs

1. Les Comités

Une des premières étapes de la mise en place d'Epidemium a été de fonder un Comité d'éthique indépendant, puis un Comité scientifique. Nécessaires dans toute recherche, ils étaient devenus indispensables par l'association des termes santé, *open* et big data au sein d'un même programme. Ces comités ont été pensés de telle façon qu'ils recouvraient par leur composition tous les secteurs concernés et toutes les expertises nécessaires à sa bonne conduite *(voir **fiche n°1a** et **fiche n°1b**)*. Leur création répondait à plusieurs volontés : garantir le

> *« Epidemium a été une aventure protéiforme merveilleuse : enrichissante sur le plan scientifique bien sûr, mais aussi sur le plan humain. Grâce à un encadrement au dynamisme constant, nous avons pu mettre à profit les meetups, RAMPs, et rencontres à La Paillasse pour fédérer des profils, compétences et personnalités variés dans un même but : la recherche contre le cancer. »*
>
> ***Benjamin Schannes, Porteur du projet Approches prédictives et risques de cancer***

sérieux et la faisabilité d'une telle initiative, accompagner une dynamique ouverte et lui permettre de se développer, innover tout en respectant une approche méthodologique et un cadre éthique rigoureux.

En amont du Challenge, le Comité d'éthique indépendant a délimité, par la Charte Epidemium *(voir **fiche n°3a**, page 137)*, le champ des possibles afin d'assurer le bon respect des usages des données au sein du programme. Il a notamment validé la faisabilité de ce dernier en s'intéressant à la captation et à la sélection des bases de données mises à la disposition des participants, l'utilisation de données nécessitant des réflexions autour des enjeux de vie privée, d'anonymisation et de consentement afin d'être bénéfique aux patients. Le Comité scientifique, quant à lui, s'est assuré de la qualité des productions de la communauté, à laquelle aucun diplôme ou certificat n'était demandé *a priori*. Il a garanti la méthodologie employée par les organisateurs du programme et par les participants. Durant le Challenge, les comités ont veillé au respect des règles par chaque projet, défini une grille de critères d'évaluation, accompagné les projets dans la formulation de leurs hypothèses et dans leur finalisation, identifié leurs applications et implications possibles, pour finalement se constituer en jury et les évaluer. Plus largement, ils ont permis d'engager une réflexion sur les pratiques actuelles, l'apport des technologies et l'utilisation des données ouvertes, notamment dans la recherche.

2. Un écosystème et des partenaires

Pour ne pas rester un vain effort, le programme Epidemium a souhaité créer un écosystème où des professionnels de la santé et de la recherche médicale, de la *data science* et du monde de l'innovation ouverte se rencontrent et échangent leurs connaissances ainsi que leurs ressources et méthodes de travail. Cette dynamique a permis à la fois de sonder l'état d'esprit des acteurs dits classiques vis-à-vis d'une telle initiative et d'interroger leur possible collaboration pour nous aider à concevoir et à délimiter le programme, et pour y participer. Ces acteurs, individus et structures, sollicités directement ou rencontrés lors de diverses présentations d'Epidemium, ont joué différents rôles. Ils ont fait connaître le programme au sein de leur écosystème en acceptant d'être des relais communautaires ou des ambassadeurs. Motivés et convaincus, ils ont également crédibilisé la démarche du programme, défendu ses valeurs et fait grandir la communauté dans les sphères et groupes les plus pertinents. Enfin, ils ont mis à la disposition de la communauté leurs expertises, ressources et outils, indispensables à la réussite du programme. Ils ont par exemple participé à l'animation de la communauté et à la production de connaissances, partagées ensuite de façon ouverte, en intervenant lors des conférences publiques du Challenge4Debate (C4D). L'expertise de cet écosystème a également bénéficié au programme, qui est alors devenu un objet de réflexion, que nous cherchions toujours à adapter aux besoins perçus des participants.

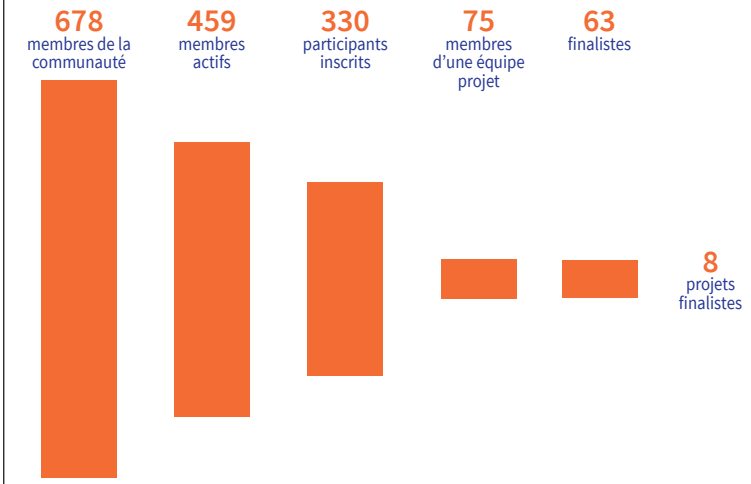
Des partenaires techniques se sont également ajoutés à cet écosystème, qui ont permis au Challenge4Cancer d'exister en ouvrant leurs outils aux participants et en acceptant de les aider dans leurs travaux. Cet engagement a permis de développer un Challenge de qualité, capable de produire des études sérieuses du point de vue du traitement des données et de la méthodologie employée. Cet environnement technique de travail, normalement réservé à des chercheurs professionnels et à des entreprises, a été optimal pour les participants. Un ensemble de partenariats techniques ont été tissés avec différentes structures : Hypercube, habituée à collaborer avec

des acteurs sur des travaux de recherche, qui développe une technologie unique d'analyse big data permettant d'explorer de façon exhaustive des phénomènes dont les causes sont complexes à comprendre et à prédire ; Dataiku, qui a développé un studio de data-analyse et de *dataviz*, mettant à disposition une large gamme d'outils *click-and-go* pour forger les intuitions et construire des hypothèses autour de *datasets* ; Teralab, un cluster big data conçu pour apporter une réponse immédiate aux besoins des chercheurs, enseignants et entreprises pour développer la connaissance et les innovations en analytique big data.

3. Une communauté transdisciplinaire et dynamique

La communauté d'Epidemium est sa véritable richesse, hétérogène tant au niveau des compétences que des profils. Cette communauté, entendue dans son ensemble, comprend

Figure n°1 : Une dynamique communautaire efficace



Description : *Taux de conversion de l'engagement communautaire.*

Source : statistiques Meetup et Epidemium.cc

« Si c'était à refaire, je participerais sans hésiter au Challenge4Cancer car cela a été très formateur pour moi, et parce que c'est une nouvelle manière de faire avancer la recherche. »

Muriel Londres
Membre du Comité d'éthique indépendant

Allier cancer et big data grâce à une méthodologie agile

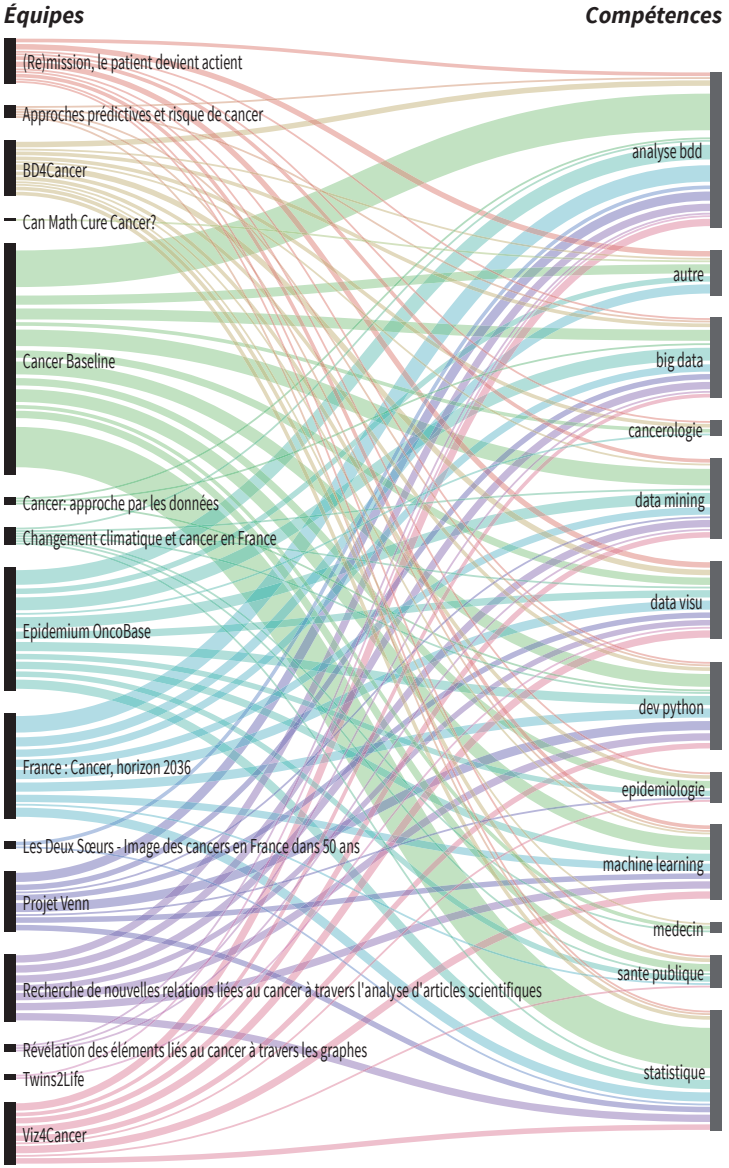
tous les acteurs qui ont contribué à un moment ou un autre à Epidemium, incluant les membres des comités, les acteurs de son écosystème et les participants déclarés du Challenge. Pour en dresser un schéma global, cela représente un peu moins de 700 membres, 330 participants inscrits, 75 personnes qui ont pris part à un projet, 63 finalistes pour 8 projets sélectionnés en finale *(voir **figure n°1**, page 23)*. Ces chiffres soulignent à la fois l'intérêt que l'initiative a suscité et l'engagement qu'elle a su entretenir tout au long des six mois.

À travers le Challenge4Cancer, l'équipe Epidemium a su faire se rencontrer et travailler ensemble des profils différents provenant principalement de trois domaines : la *data science*, l'informatique et la santé. Cela correspond aux spécificités ainsi qu'aux besoins du Challenge et de ses thématiques. Si la communauté s'est d'abord portée sur l'univers de la *data science*, cela est certainement dû au fait que le C4C était fondé sur les données, et par conséquent était plus proche de la culture des *data scientists*. Le programme a donc également dû être pensé comme un lieu d'acculturation à ce format et à ces techniques, notamment pour les médecins.

Pour approfondir cette typologie, il est à noter qu'il y a eu 1 176 compétences cumulées lors du Challenge. En étudiant les projets et les compétences que ceux-ci ont déclaré avoir utilisées, nous pouvons observer une véritable circulation de ces dernières. Sur les 15 projets enregistrés, tous ont fait intervenir à un moment ou un autre de leur développement plusieurs des onze compétences clefs *(voir **figure n°2**, page ci-contre)*. Cela permet d'identifier la logique d'émulation favorisée par le Challenge, par les outils ainsi que par les nombreux moments de rencontre et d'échange.

Enfin, nous pouvons constater que la communauté s'est accrue tout au long des six mois, et ce, dans chaque type de compétences. Cela souligne l'intérêt qu'a su susciter le Challenge et sa capacité à convaincre les curieux. Comme nous l'avons remarqué plus haut, l'arrivée tardive des acteurs de la santé peut s'expliquer par sa nature, qui nécessitait à son début un important travail de nettoyage et d'agrégation des données. Les professionnels de santé ont participé principalement lors

Figure n°2 : La transdisciplinarité des équipes

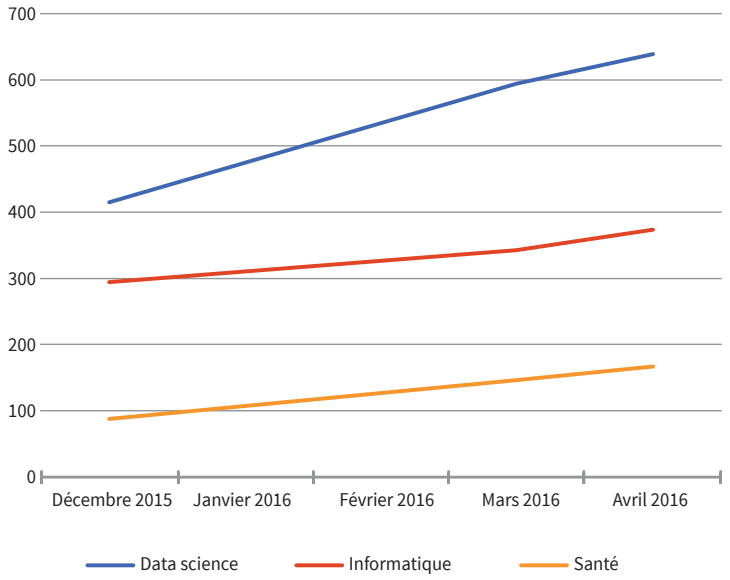


Description : *La répartition des compétences au sein des quinze équipes-projets.*

Source : Epidemium, visualisation réalisée avec <http://raw.densitydesign.org/>

de la seconde moitié du Challenge *(voir **figure n°3**, ci-dessous)*, au moment où leur expertise était demandée afin d'interpréter les premiers résultats obtenus et de guider les hypothèses et les solutions proposées sur le plan médical.

Figure n°3 : Domaines de compétences et leur répartition temporelle lors du C4C



Description : *Évolution des compétences déclarées sur la plateforme Epidemium.cc lors du Challenge4Cancer.*

Source : Epidemium.cc

Toutefois, agréger une communauté ne suffit évidemment pas ; il faut ensuite qu'elle vive. Ce groupe de personnes, s'il est dévoué à une cause commune, n'en a pas moins besoin d'être animé au jour le jour pour rester mobilisé, d'autant plus que son engagement est bénévole. Nous avons donc travaillé à créer une dynamique communautaire forte, et ce, grâce à des ressources variées.

Des ressources pour structurer, développer et animer le programme

Pour faire connaître le programme, engager puis animer la communauté et lui donner les moyens de mener à bien ses projets, l'équipe coordinatrice a mis en place plusieurs outils, à la fois *online* et *offline*. Chacun d'eux a eu un rôle spécifique et répondait à un besoin qui fut soit pensé en amont soit exprimé pendant le Challenge par les participants *(voir **fiche n°1d** : La boîte à outils d'Epidemium, page 62)*.

Si une hiérarchie des outils devait être définie, les outils *online* centraux du Challenge et structurants pour la communauté seraient le site et le Wiki[2]. Le site web[3], premier outil pensé et conçu en amont du Challenge, avait pour fonction de répondre à la logique communautaire. Porte d'entrée pour participer au Challenge4Cancer, il présentait le programme mais surtout permettait de gérer la communauté et de la rendre intelligible. Pour ce faire, il catégorisait les participants selon plusieurs critères : les profils, les compétences, l'appartenance ou non à un projet et le fait de concourir dans telle ou telle thématique. À partir de ces informations, un moteur de recherche offrait alors de naviguer dans la communauté virtuellement recensée et qualifiée. Il était par exemple possible de voir les projets existants dans chacune des thématiques, de rechercher des compétences et des profils particuliers pour mener à bien un projet, d'identifier les porteurs de projet afin de les contacter, *etc*. Ainsi, le site avait pour objectif premier de mettre les compétences en interaction et de favoriser leur identification pour ceux qui souhaitaient participer ou développer leur projet.

Le Wiki a quant à lui occupé une place prédominante du fait de sa grande modularité. L'équipe coordinatrice a pu l'adapter aux différentes étapes du programme, à ses besoins mais aussi à ceux des participants. Cela est constatable en quelques chiffres : selon les statistiques fournies par le Wiki d'Epidemium, de sa création jusqu'à la fin du programme, il y a eu 3 136 modifications, 10 024 pages vues et 3 276 contributions pour 117 contributeurs. Outil quotidien librement accessible et modifiable, il est devenu le véritable pouls du programme. Il

> *Cancer & big data : la science collaborative s'organise avec Epidemium »*
> ***Hugo Jalinière***
> *(Sciences et Avenir, 30/05/2015)*

Allier cancer et big data grâce à une méthodologie agile

a tout d'abord été un moyen de communication, centralisant l'ensemble des informations nécessaires à la compréhension du programme et du Challenge4Cancer ainsi que toutes les actualités : présentation des initiateurs et des dimensions d'Epidemium, des partenaires, des comités et de leurs réflexions, des modalités de participation, des événements ponctuels ou récurrents, *etc*. Le Wiki centralisait également les liens vers les autres outils en expliquant leurs fonctions et leurs modalités d'accès. Enfin, et c'est par cette caractéristique qu'il a été utilisé à la fois par l'équipe Epidemium et par les participants, il a répondu à une logique de documentation et, par conséquent, aux impératifs de transparence et d'ouverture. La totalité du projet a été documentée, notamment grâce à la mise en place d'un Carnet de Bord [4] retranscrivant chaque semaine les faits marquants du programme et des projets, mais aussi à travers les comptes-rendus des événements. Cette dimension va de pair avec la volonté de produire des savoirs accessibles, que chacun peut se réapproprier librement et gratuitement. Les participants avaient en outre la possibilité de réagir, via des fenêtres de discussion, aux différents contenus. Enfin, mis entre les mains des participants, ce fut l'endroit où ils durent présenter et documenter leur projet et leurs hypothèses. Ainsi, à la fin du Challenge, lors du gel des pages projets nécessaire pour qu'elles soient évaluées par le jury, celles-ci regroupaient les informations suivantes : l'objectif du projet, la présentation de l'équipe, de la production finale, des ressources utilisées (jeux de données, outils, *etc*.) et de la méthodologie employée, et les développements futurs imaginés.

Ensuite, Epidemium a mis en place plusieurs outils satellites dédiés à des problématiques communautaires ciblées, notamment liées à la vie du programme. Une plateforme, sous forme de questions-réponses [5], permettait aux membres de la communauté de poser des questions liées aux thématiques, à l'utilisation des données en cancérologie et aux méthodes employées lors du Challenge, auxquelles participants, experts et équipe coordinatrice pouvaient répondre librement. À travers le groupe Epidemium sur Facebook ainsi que le compte Twitter, l'équipe coordinatrice pouvait échanger avec la communauté

plus étendue sur l'actualité du programme, du Challenge et celle des projets. Le premier étant davantage à destination de la communauté Epidemium et le second à destination d'un écosystème plus large. Ces deux outils ont été le moteur d'une activité importante de curation scientifique collective autour des thématiques du cancer et du big data. Sans oublier un compte sur la plateforme Meetup [6], pour soutenir et faire connaître l'activité événementielle du programme.

Enfin, des outils purement techniques furent également nécessaires à Epidemium et aux équipes pour traiter les données. La plateforme Epidemium Portail *Open Data* [7], utilisant le logiciel *open source* CKAN, répondait à cet enjeu technique en rendant accessibles plus de 21 000 jeux de données pour le Challenge4Cancer via un moteur d'exploration. Elle les recense selon qu'ils sont liés à la démographie, à l'environnement et à l'agriculture, au travail, à l'économie, au comportement des individus, à la santé et au cancer, rendant ainsi intelligible la masse de données, grâce à une première grille de lecture facilitant leur appréhension. De plus, comme nous l'avons vu, plusieurs outils d'analyse furent mis à la disposition des participants grâce aux partenaires du programme : un cluster big data par Teralab, un outil de data analyse par HyperCube et un studio de data analyse et de dataviz par Dataiku.

Loin d'être un programme dématérialisé, Epidemium s'est incarné en de nombreux événements *(voir* ***fiche n°1e*** *: Call4Debate 2015-2016, page 63)* visant à faire se rencontrer les participants, à favoriser les synergies et à créer des moments d'échange avec des experts. Ces événements communautaires, vingt-et-un au total, se sont déclinés en différents formats selon leurs objectifs et les publics ciblés.

Le plus répliqué a été celui des conférences, où sont intervenus plusieurs experts pour présenter des cas concrets issus de leurs travaux actuels ou passés, puis pour échanger avec le public à propos de problématiques connexes. Ces conférences informelles avaient pour rôle d'aider les participants dans leurs réflexions et dans la réalisation de leur projet et, étant retranscrites sur le Wiki, de permettre d'engendrer de la connaissance à destination de la communauté. Ces événements

Allier cancer et big data grâce à une méthodologie agile

étaient ouverts à tous et gratuits afin de faire connaître le programme et d'attirer un large public intéressé par les sujets abordés. Au total, 926 personnes ont assisté aux conférences.

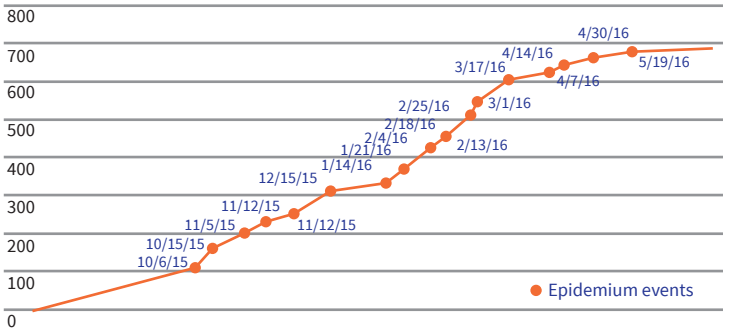
Le deuxième format était celui des RAMPs (*Rapid Analytics and Model Prototyping*) [8], c'est-à-dire des *data challenges* fondés sur l'idée de coopétition, propre à l'état d'esprit du programme, l'ensemble des productions des participants étant visibles par tous. Ces événements d'une journée, plutôt à destination des *data scientists*, également ouverts et gratuits, ont eu pour objectif de créer et de tester des modèles de prédiction en se basant sur les données récoltées par Epidemium. Les participants, constitués ou non en équipe, présents dans la même salle, soumettaient leurs modèles sur un serveur de façon ouverte et leur performance était affichée sur un *scoring board*. Tous avaient accès librement aux codes des modèles déjà soumis et ainsi pouvaient comprendre ce qui était efficace ou non, reprendre des éléments, les combiner, les améliorer pour soumettre à leur tour un code nouveau. Cette approche hybride devait aboutir, en un temps rapide, à un modèle de prédiction performant qui résultait des efforts de tous.

Le troisième format, les événements dits *Bocal*, pensés comme plus restreints, étaient des rendez-vous ponctuels visant à faire se rencontrer les participants du Challenge, pour d'abord faciliter la constitution des équipes, puis répondre à leurs besoins précis en la présence d'experts, notamment médicaux.

Enfin, d'autres événements visant à rythmer la vie du programme et à insuffler une dynamique aux six mois de Challenge ont été organisés : la soirée de lancement qui a constitué la première rencontre entre les membres de l'écosystème (partenaires, membres des comités et participants), le point de mi-parcours où les équipes étaient invitées à présenter leur projet, méthode et hypothèses aux comités afin que ceux-ci puissent les guider dans leurs réflexions, et la soirée de clôture qui a constitué la finale du Challenge, durant laquelle les équipes pouvaient pitcher leur projet devant la communauté et le jury, qui délibérait ensuite. Ces moments ponctuant les étapes du Challenge ont posé des échelons pour les participants et nous ont donc aidés à entretenir la dynamique de travail communautaire.

Ces différentes rencontres ont permis de répondre à la problématique du temps long, de l'engagement fluctuant et du bénévolat. Avec cette stratégie événementielle, Epidemium a souhaité créer et maintenir l'intérêt des participants en leur proposant des moments à la fois ludiques, utiles et formateurs. Dépassant nos espérances, cette partie a constitué un réel atout pour le programme comme le souligne la corrélation entre les événements et la croissance de la communauté *(voir **figure n°4**, ci-dessous)*.

Figure n°4 : Corrélation entre la croissance de la communauté et les événements



Description : *Évolution de la communauté au regard des événements organisés par Epidemium.*

Source : Meetup statistics

Grâce aux outils mis en place, Epidemium se définit comme un format agile, capable d'anticiper, de percevoir et de répondre aux besoins des participants. Tous ces outils ont facilité une constante relation entre la communauté, les experts et l'équipe Epidemium. Enfin et surtout, on constate une interaction entre les formats *online* et *offline*, chaque dimension favorisant le succès de l'autre.

« Nous sommes très fiers d'être partenaire technologique d'Epidemium sur ce premier challenge. En mettant à disposition du consortium notre plateforme collaborative d'analyse prédictive, nous avons eu la chance de suivre et d'accompagner de très beaux projets autour de la recherche sur le cancer. Passionnant et inspirant. »

Thomas Thus, Dataiku

Les succès et insuccès de la méthodologie

Le fait même de développer Epidemium sous la forme d'un programme d'*open science* en cancérologie fondé sur le big data, a contribué à sa réussite. Ces problématiques contemporaines, médiatisées et évocatrices ont pu éveiller la curiosité et faciliter le développement de l'écosystème et de la communauté. Bénéficiant de la vague médiatique associée aux thématiques d'ouverture et de big data, Epidemium a su tirer partie de l'actualité pour mobiliser et réunir les compétences nécessaires, ainsi que pour montrer la faisabilité des nouvelles approches dans ce domaine encore peu fourni en réalisations concrètes. Ce fut alors l'occasion de se positionner à la pointe d'un sujet en y prenant part activement par le Challenge, tout en le questionnant grâce à l'intervention et à la participation d'experts ; sujet qui s'impose de plus en plus comme un enjeu clef pour l'avenir.

Le Challenge4Cancer n'a eu aucune barrière à l'entrée : aucune accréditation n'était demandée, les moyens techniques ont été fournis, le programme a développé un cadre et un écosystème sécurisant, formateur pour tous les participants, même les plus novices. Bien plus, l'ouverture choisie comme méthode a structuré le Challenge, respectant le choix d'une transparence et d'une intelligibilité totale et pour tous, notamment par un effort de documentation, ce qui a favorisé son accessibilité. De plus, travailler sur des données déjà ouvertes et librement accessibles a rendu le programme faisable. Juridiquement, cela a facilité sa mise en place puisque les données choisies, déjà disponibles et accessibles librement, répondaient au traitement légal exigé par la loi française. De plus, ces données ouvertes étaient assez conséquentes et hétérogènes pour constituer un matériau de recherche riche et prometteur et, étant accessibles à tous sans discrimination, elles constituaient un commun dont il était légitime de s'emparer. Epidemium propose de redonner sens de façon collaborative à nos propres données et même de se les réapproprier.

Au-delà de cet aspect d'ouverture omniprésent et de la volonté de chacun des acteurs impliqués de faire partie d'une cause commune, portés par un véritable intérêt scientifique et humain, le programme a été le réceptacle et le catalyseur de nombreuses

motivations aussi diverses que ses différentes parties prenantes. En dépit de l'hétérogénéité des profils des participants, *data scientists*, médecins, employés, étudiants, chercheurs d'emploi, *etc.*, il est possible de souligner des motivations, partagées par tous, qui ont pu inciter les individus à participer au Challenge.

Ce fut l'occasion de découvrir et de se former à des sujets nouveaux, de développer de nouvelles compétences mais aussi d'approfondir celles déjà acquises en les mettant au service d'une cause commune, de les tester dans le cadre d'un programme concret, de participer à une expérience collaborative et de rencontrer des personnes provenant de secteurs différents, notamment du domaine de l'*open*, de la santé et de la *data science*. Preuve en est l'enthousiasme que le programme a suscité chez les étudiants, qui ont constitué environ le quart des participants.

La première édition d'Epidemium, qui se clôt véritablement avec ce *Livre blanc*, nous a donc permis de tester la viabilité d'une démarche communautaire et ouverte pour appréhender le cancer différemment, ainsi que de mesurer l'intérêt grandissant que ce projet a suscité à la fois chez les acteurs accrédités de la recherche médicale et du big data, chez les institutions publiques et chez les nombreux individus qui se reconnus dans nos valeurs et notre mission.

Dans la continuité de notre engagement, nous nous ferons guider par les retours de notre écosystème et par les enseignements tirés de l'expérience du travail communautaire que nous avons eu le plaisir de coordonner. ■

1. Coopétition : néologisme fondé sur l'association des termes compétition et collaboration, propre à l'état d'esprit du programme.
2. Wiki d'Epidemium <http://wiki.epidemium.cc/wiki/Accueil>.
3. Site web Epidemium <http://epidemium.cc>.
4. Carnet de Bord Epidemium <http://wiki.epidemium.cc/wiki/Carnet_de_bord>.
5. Epidemium Q&A <http://qa.epidemium.cc/>.
6. Epidemium Meetup <http://www.meetup.com/fr-FR/Epidemium/>.
7. Epidemium portail Open Data <http://data.epidemium.cc/dataset>.
8. Le RAMP (*Rapid Analytics et Model Prototyping*) est un outil, développé par le Paris-Saclay Center for Data Science et l'Ecole des Mines, pour la gestion des data challenges, <http://www.ramp.studio/>.

Le pouls du programme

SCIENCE DES DONNÉES | RÉSEAU SOCIAL | COLLABORATION | ANALYTIQUE | DONNÉE NUMÉRIQUE

Grâce aux nouvelles technologies, il est aujourd'hui possible d'analyser quantitativement l'activité d'équipes travaillant de manière collaborative. Dans le cas du Challenge4Cancer organisé par Epidemium, plusieurs outils numériques ont été mis à la disposition des participants. Ainsi, c'est l'analyse de l'activité de ces outils utilisés par tous, comme le site web, le Wiki, le Q&A et la plateforme Meetup qui nous permet de cerner les comportements des membres de la communauté au sein du Challenge, de mesurer leur engagement et finalement de formuler des recommandations pour la suite du programme.

// AUTEURS

Djalel Benbouzid / Léo Blondel / Marc Santolini

L'équipe de coordination d'Epidemium a mis en place plusieurs outils en ligne lors du Challenge4Cancer (C4C) afin de créer un environnement de travail virtuel collaboratif dans lequel les membres de la communauté pouvaient s'informer sur le programme, le Challenge et les thématiques connexes, mais aussi interagir entre eux et mener à bien leur projet.

Notre analyse se fonde sur les données collectées sur le groupe Meetup d'Epidemium <www.meetup.com/fr-FR/Epidemium> ainsi que sur les trois outils mis en place par l'équipe Epidemium et largement utilisés par tous les participants au C4C :

1. Le site web <http://epidemium.cc>, ouvert le 5 novembre 2015, qui permettait principalement de s'inscrire au C4C et s'informer ;
2. Une partie Wiki <http://wiki.epidemium.cc>, ouverte le 1er octobre 2015 et rendue publique le 5 novembre 2015, pour documenter et partager avec la communauté les avancées des différents projets ;
3. Une section Q&A <http://qa.epidemium.cc>, ouverte le 23 février, pour échanger avec tous les membres de la communauté élargie en posant des questions et en répondant à celles d'autres membres.

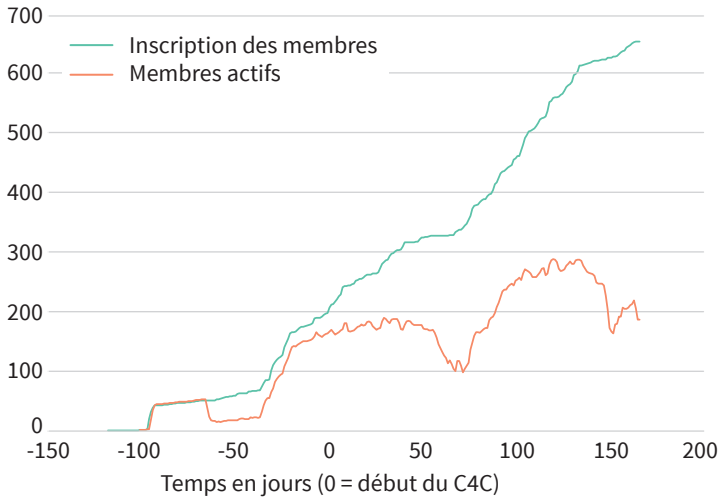
Toutes les données collectées pour la rédaction de cet article, ainsi que les formules employées pour leur analyse sont disponibles sur GitHub <https://github.com/Epidemium/LivreBlanc>.

Une communauté engagée

Le premier niveau d'analyse de données est celui des individus face aux outils en ligne mis à leur disposition par le programme lors du Challenge4Cancer. À partir de cette première lecture, nous pouvons avoir un aperçu assez précis des comportements des membres de la communauté et mesurer leur niveau d'engagement et d'appropriation face à ces différents outils en ligne.

Tout d'abord, dans la ***figure n°1*** *(voir page 36)*, il est possible d'observer l'évolution temporelle de l'activité des membres sur le groupe Meetup d'Epidemium depuis sa création, le

Figure n°1 : Analyse temporelle de l'activité des membres - données agrégées



Description : *Évolution temporelle du nombre de membres sur la plateforme Meetup. La courbe verte correspond au nombre total de membres inscrits, et la courbe orange au nombre de membres ayant eu une activité sur la plateforme au cours des 30 jours précédents.*

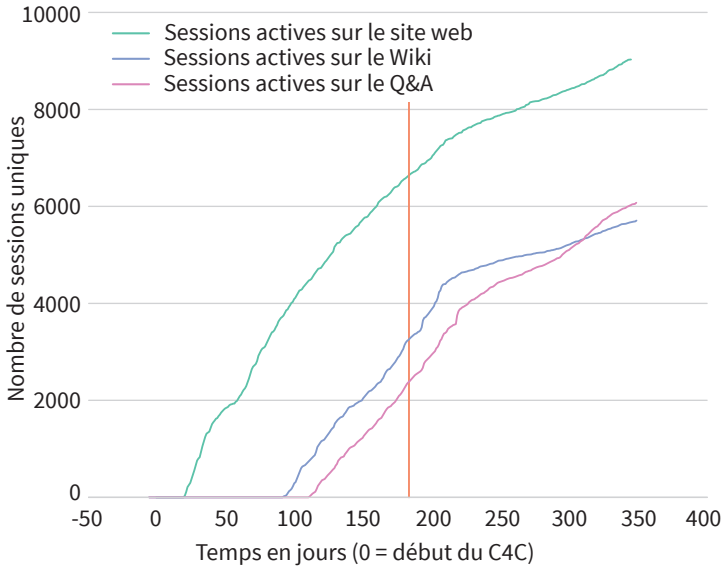
Source : Meetup.com, groupe Epidemium ; graphique réalisé par les auteurs

9 septembre 2015, à sa clôture, le 5 mai 2016. Le nombre d'utilisateurs inscrits, ici en vert, a connu une augmentation stable tout au long du Challenge, avec cependant un creux correspondant à la période des vacances de Noël. Avec le nombre de membres actifs en orange, c'est-à-dire qui ont visité le groupe au moins une fois durant les trente jours précédents, nous pouvons voir émerger plus clairement deux périodes. La première commence peu avant le début du Challenge, puis la seconde démarre après le creux d'activité des fêtes de fin d'année, et voit une augmentation d'environ 50% sur l'activité des membres. Cette tendance a été renforcée par « le point de mi-parcours », événement organisé le 12 mars 2016, pour les équipes qui souhaitaient y participer, et qui leur proposait de

venir confronter leurs approches, leurs hypothèses et leurs méthodologies aux membres des deux comités d'éthique et scientifique, sur la base d'une documentation intermédiaire, d'une part, et d'une présentation orale publique de leurs avancées, d'autre part.

La ***figure n°1*** montre ainsi qu'il y a eu un recrutement constant de nouveaux membres pour participer au Challenge4Cancer. Par ailleurs, la courbe des nouvelles inscriptions ne montre aucun signe de saturation et, par conséquent, laisse à penser que le nombre potentiel d'individus à la fois intéressés et susceptibles de participer au Challenge était bien supérieur au nombre réellement atteint.

Figure n°2 : Analyse temporelle de l'activité des membres sur les trois outils *online*



Description : *Évolution temporelle du nombre de sessions ouvertes sur les trois outils mis à disposition par Epidemium aux membres de la communauté (site web, Wiki et Q&A). La ligne rouge correspond à la fin du C4C.*

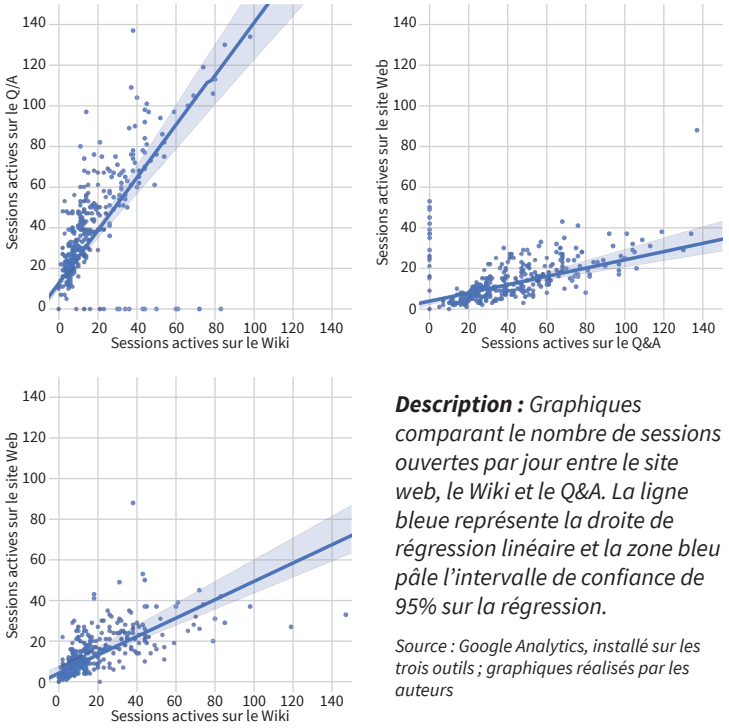
Source : Google Analytics, installé sur les trois outils ; graphique réalisé par les auteurs

Le pouls du programme

Avec la ***figure n°2*** *(voir page 37)*, nous pouvons observer l'évolution temporelle des visites des trois outils en ligne mis à la disposition des membres (site web, Wiki et Q&A) pour en développer une analyse à partir du lancement officiel du C4C jusqu'au 20 juillet 2016. Le décalage des courbes tient du fait que les données *Google analytics* ont été récupérées sur des périodes de temps différentes. En effet, la collection de données a commencé le 14 novembre 2015 pour le site web, le 2 février 2016 pour le Wiki et le 23 février 2016 pour le Q&A. On note que l'attraction des trois plateformes a été stable et similaire (pentes semblables) au cours du Challenge, avec une nette réduction des visites suite à l'événement de clôture et la finale du C4C. Néanmoins, les visites ont continué jusqu'à trois semaines après la fin du Challenge, révélant ainsi un intérêt de la part de la communauté mais aussi certainement d'acteurs extérieurs pour la lecture des résultats du Challenge.

Enfin, les panels de la ***figure n°3*** *(voir ci-contre)* montrent les corrélations d'utilisation entre les trois outils durant le Challenge4Cancer, de son lancement, le 5 novembre 2015, jusqu'au lendemain de sa clôture, le 6 mai 2016. Lorsqu'un utilisateur se connecte, une session est ouverte sur Google analytics, permettant de suivre les utilisateurs sur les trois outils : il est donc possible de savoir quelles plateformes ont été visitées. Chaque jour, le nombre total de visites par site est recueilli, chaque utilisateur comptant de manière unique quelque soit le nombre de visites effectuées. Le nombre de visites des différentes plateformes est ensuite comparé. On observe une très forte corrélation entre les visites sur les trois outils. Ainsi, nous pouvons en déduire que les visiteurs et participants du Challenge ont utilisé de manière égale l'ensemble des outils mis à leur disposition et que ces derniers, loin d'être redondants, s'inscrivent dans une réelle complémentarité et répondent chacun à un besoin des participants, qu'il ait été anticipé ou révélé lors du Challenge.

Figure n°3 : Corrélation entre le site web, le Wiki et le Q&A



Description : *Graphiques comparant le nombre de sessions ouvertes par jour entre le site web, le Wiki et le Q&A. La ligne bleue représente la droite de régression linéaire et la zone bleu pâle l'intervalle de confiance de 95% sur la régression.*

Source : Google Analytics, installé sur les trois outils ; graphiques réalisés par les auteurs

Constitution des équipes et méthodes de travail

En revanche, c'est la nature même du programme, qui cherche à appréhender l'épidémiologie du cancer différemment en exploitant le potentiel du big data, qui a stimulé la collaboration entre experts de domaines habituellement très éloignés.

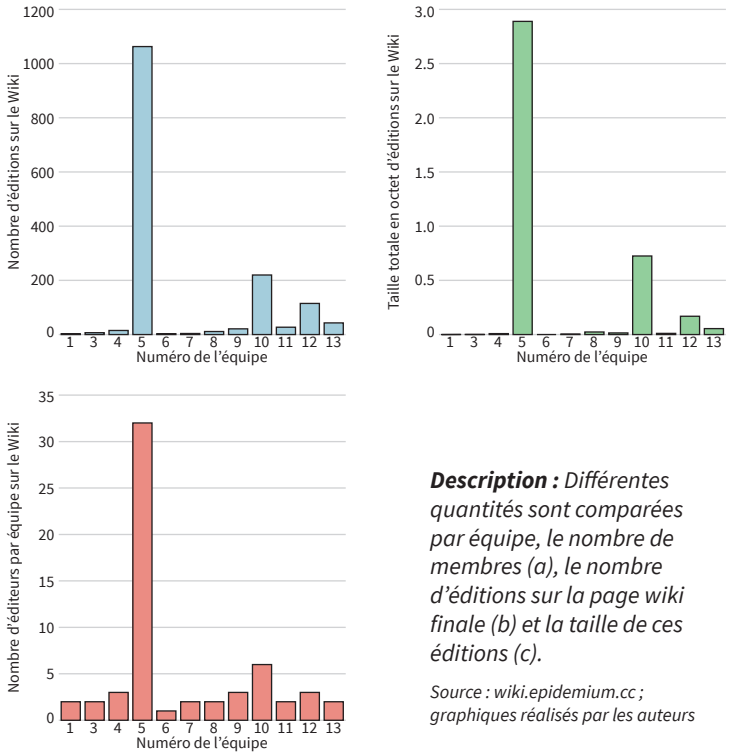
Nous présentons ici cette dimension collaborative en y dévoilant les aspects de dynamique temporelle et de structure interne.

Regroupement par équipes

De la lecture de ces trois graphiques émerge nettement le constat qu'une équipe a mobilisé un nombre bien plus important de contributeurs que les autres. En analysant le contenu des pages Wiki qui, dans l'ergonomie du Challenge, recueillent le travail des équipes, nous observons que quatre d'entre elles ont produit la majorité des « éditions » ou *edits*, que ce soit en nombre ou en taille.

Ainsi, le format du Challenge a conduit à un regroupement des forces de travail en quelques équipes productives plutôt qu'à une multiplication des équipes indépendantes les unes par

Figure n°4 : Analyse de l'édition du Wiki par différentes équipes



Description : *Différentes quantités sont comparées par équipe, le nombre de membres (a), le nombre d'éditions sur la page wiki finale (b) et la taille de ces éditions (c).*

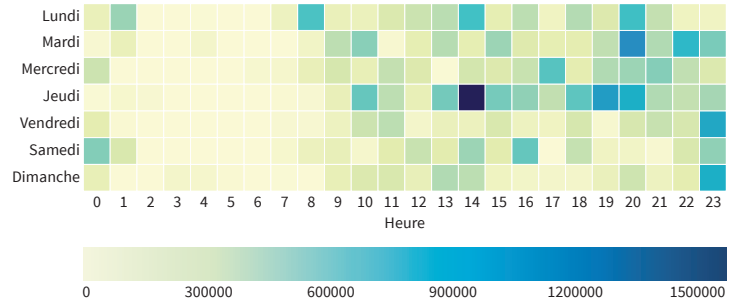
Source : wiki.epidemium.cc ; graphiques réalisés par les auteurs

rapport aux autres. Il faut noter que la productivité est mesurée ici par l'édition Wiki, ce qui ne représente pas la totalité du travail fourni : certaines équipes ont pu, par exemple, produire du code sur Github, créer des notebooks Jupyter[1] ou utiliser d'autres outils, extérieurs à ceux mis à disposition directement par Epidemium. Ainsi, nos analyses dévoilent la partie émergée de l'iceberg et ne peuvent rendre compte de la productivité réelle de tous les groupes.

Analyse des compétences et temporalité du travail collaboratif

Comme nous pouvons le déduire de la ***figure n°5*** *(voir ci-dessous)* l'engagement de la communauté, mesuré par l'activité des participants au C4C, a été assez uniforme au cours de la semaine, avec un pic d'activité le jeudi midi, et de manière générale dans l'après-midi et en soirée, donc notamment sur leur temps libre et les pauses repas. En regardant les interactions entre les différentes équipes dans la ***figure n°6*** *(voir page 42)* émerge clairement le fait que très peu de participants ont contribué à plusieurs projets, même si cela n'était pas interdit par le Règlement. Cette concentration des efforts individuels

Figure n°5 : Distribution de l'activité des membres au cours de la journé et de la semaine

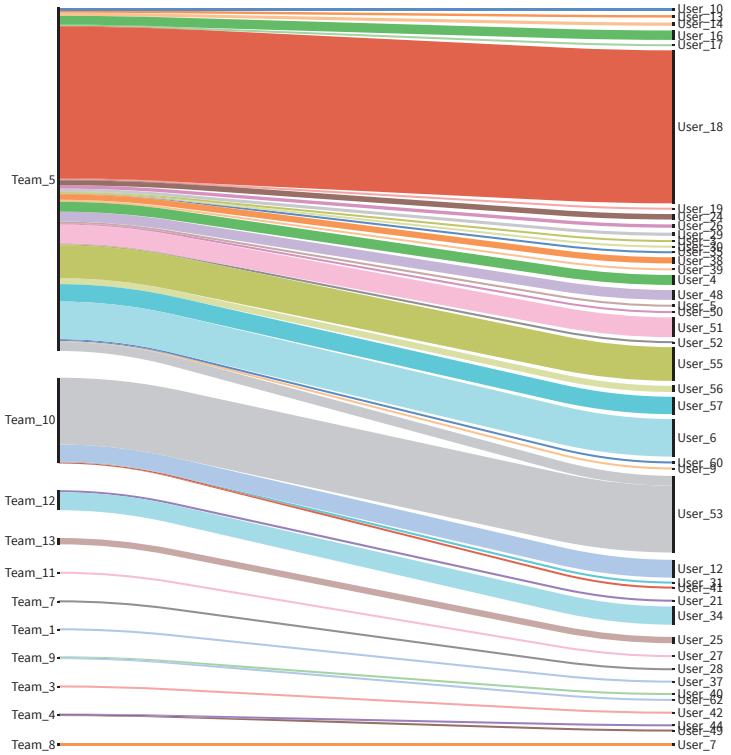


Description : *Heatmap montrant le nombre d'edits par heure et jour de la semaine.*

Source : wiki.epidemium.cc ; graphiques réalisés par les auteurs

sur des projets uniques a sûrement réduit la dispersion d'énergie mais elle a peut-être aussi limité le développement de projets transversaux à plusieurs équipes. Afin de mettre à l'échelle un tel Challenge, il sera important de penser la systématisation d'une meilleure porosité et coopération entre équipes pour articuler et aligner les projets spécifiques au bénéfice d'une vision d'ensemble.

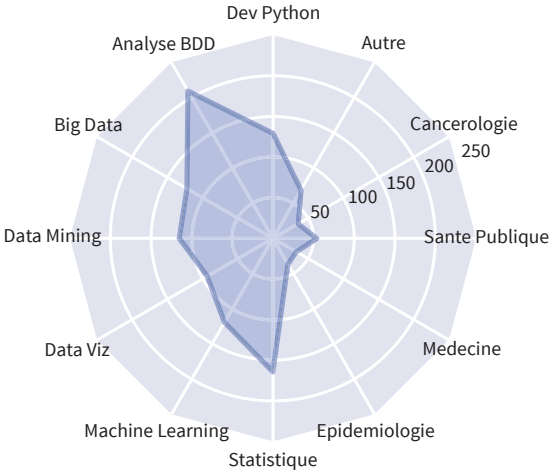
Figure n°6 : Distribution des membres au sein des équipes et leur activité



Description : *Diagramme de Sankey montrant la distribution des membres au sein des équipes et leur activité sur le Wiki (la taille d'un membre représente son nombre d'éditions Wiki).*

Source : wiki.epidemium.cc , graphiques réalisés par les auteurs

Figure n°7 : Répartition des compétences au sein des équipes



Description : *Représentation radar montrant la distribution des compétences individuelles au sein des équipes.*

Source : epidemium.cc, aptitudes déclarées par les participants lors de leur inscription ; graphiques réalisés par les auteurs

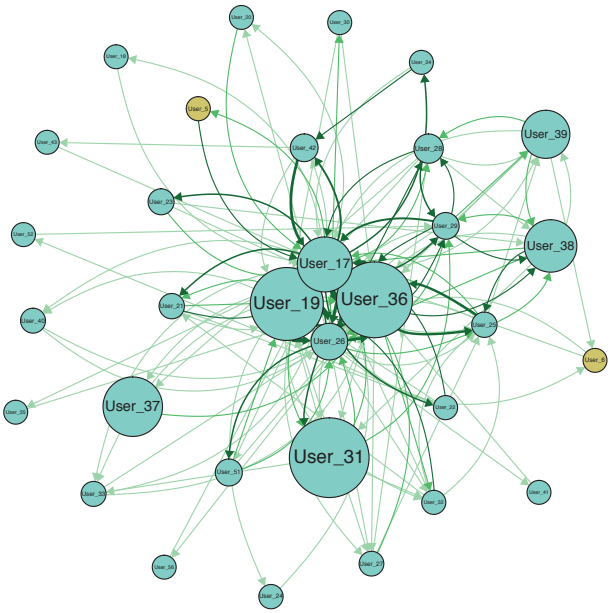
Enfin, la ***figure n°7*** *(voir ci-dessus)* nous donne un aperçu très clair de la riche pluridisciplinarité de la communauté, tout en montrant que la participation du monde de la médecine et de la santé reste certainement à renforcer en dépit d'une thématique générale du programme assez équilibrée entre médecine et big data.

Structure et dynamique de travail des équipes : analyse de l'édition du Wiki de l'équipe n°5

Nous présentons ici les résultats de l'analyse des données du Wiki de l'équipe 5, la plus active au sein du Challenge4Cancer et celle qui possède les données les plus fournies sur le Wiki.

En ***figure n°8*** *(voir ci-dessous)* figurent les réseaux de conversations issues des pages projets. Ces réseaux sont construits en joignant, par un lien dirigé, deux utilisateurs lorsque l'un a édité juste après l'autre. La direction va du prédécesseur au suiveur. La couleur plus ou moins foncée d'un lien dépend du nombre de fois qu'une telle succession a été mesurée. La taille des nœuds, quant à elle, représente le nombre d'*edits* d'un utilisateur donné. L'effet réseau est très intéressant chez l'équipe 5, pour laquelle de nombreux membres ont contribué à la page

Figure n°8 : Réseau d'interactions au sein de l'équipe 5 dans l'éditions du Wiki



Description : *Une flèche pointe d'un utilisateur A à un autre utilisateur B si B a édité le Wiki après A. Le poids des liens est marqué par une intensité de couleur proportionnelle au nombre de fois où B succède à A. La taille de chaque nœud est proportionnelle au nombre d'edits correspondant. Les nœuds en jaune correspondent aux membres de l'équipe organisatrice Epidemium.*

Source : wiki.epidemium.cc ; graphiques réalisés par les auteurs

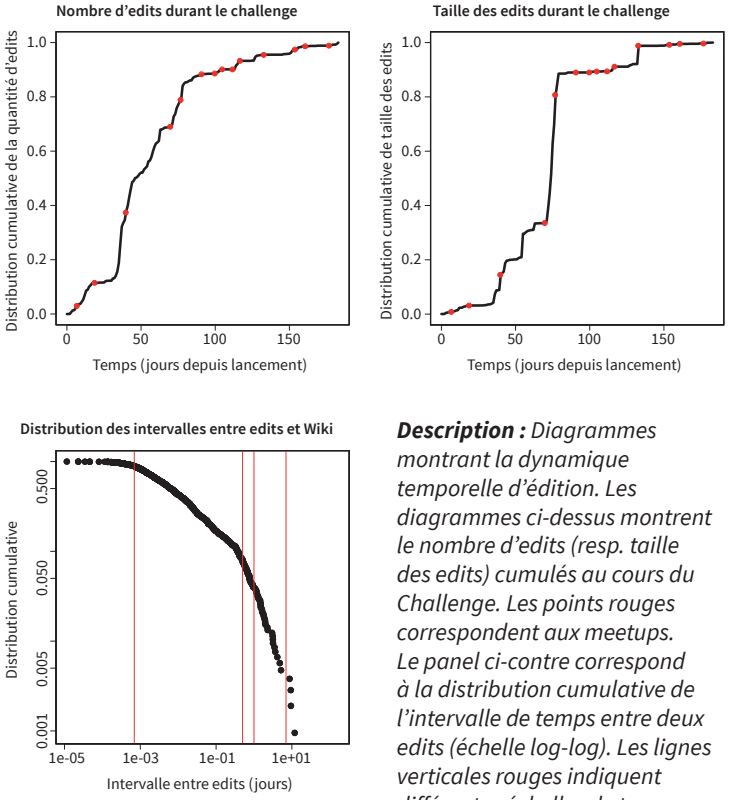
Wiki du projet, faisant apparaître un réseau dense et collaboratif. Au sein de ce dernier, un groupe d'acteurs se démarque plus particulièrement, qui semble avoir joué un rôle important dans la gestion du projet. Cet effet réseau peut être quantifié par le degré des nœuds ***(figures n°8)***. Cette mesure quantifie l'importance d'un nœud en mesurant le poids total des liens qui le relient aux autres nœuds du réseau. Cela permet de distinguer un *leadership*, si ce n'est du projet, du moins de l'écriture de la page projet.

Les ***figures n°9*** *(voir page 46)* montrent les analyses temporelles détaillées de l'édition du Wiki. Les deux premiers graphiques montrent les distributions cumulatives du nombre d'*edits* et la taille de ces derniers au cours du temps. Nous qualifions d'*edit* une soumission par un utilisateur ; sa taille dépend de la quantité de texte soumise. Les points rouges représentent les événements Meetup organisés par l'équipe Epidemium. Dans le cas de l'équipe 5, nous constatons que la majorité des *edits* a été réalisée sur une période relativement courte. De plus, les *edits* les plus importants en taille précèdent généralement un événement, ce qui peut dénoter un effet de préparation de type date limite. Enfin, le dernier graphique montre la distribution de l'intervalle de temps entre deux *edits*. Ceci informe sur la manière de travailler d'une équipe. En particulier, lorsque représentée en échelle logarithmique (*log-log*), une distribution montrant une queue linéaire est indicatrice d'une forme de travail par « salves » (*bursts*) - on parle de distribution invariante d'échelle[2]. Ainsi, alors que la plupart des éditions se suivent de près, formant des salves d'activité, il y a de manière occasionnelle des temps exceptionnellement longs entre deux éditions. Les lignes rouges indiquent les intervalles de temps correspondant à la minute, la demi-journée (12h), la journée et la semaine. On observe au sein de l'équipe 5 un tel comportement par salves marqué au cours d'une journée de travail (période avant la deuxième barre rouge) ainsi qu'au cours de la semaine, même si moins fort (période après la deuxième barre rouge, pente plus forte).

Ces résultats se généralisent aux autres équipes ayant fait assez d'*edits* sur le Wiki pour être pris en compte *(**figure n°12**,*

voir page 48). Il est d'abord à noter que l'équipe coordinatrice Epidemium doit être mise à part en tant qu'elle est l'équipe organisatrice et que son édition du Wiki donne lieu à un contenu particulier, à savoir la documentation du programme dans son ensemble et des informations quant à sa structure. Il apparaît alors une édition relativement stable au cours du Challenge et une manière de travailler par salves similaire à l'équipe 5. Cela montre l'effort continu de supervision du Wiki, permettant sans

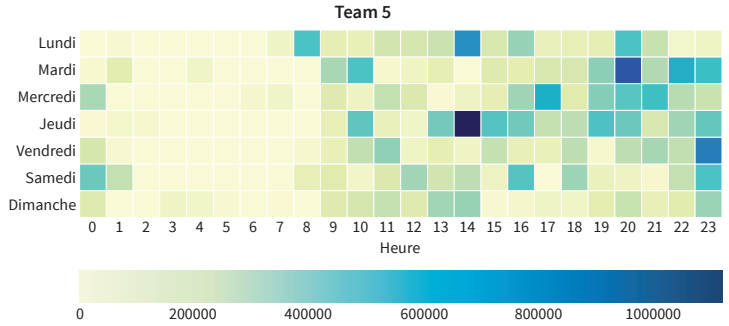
Figure n°9 : La dynamique temporelle du travail de l'équipe 5



Source : wiki.epidemium.cc ; graphiques réalisés par les auteurs

Description : *Diagrammes montrant la dynamique temporelle d'édition. Les diagrammes ci-dessus montrent le nombre d'edits (resp. taille des edits) cumulés au cours du Challenge. Les points rouges correspondent aux meetups. Le panel ci-contre correspond à la distribution cumulative de l'intervalle de temps entre deux edits (échelle log-log). Les lignes verticales rouges indiquent différentes échelles de temps : minute, demi-journée (12h), jour, semaine.*

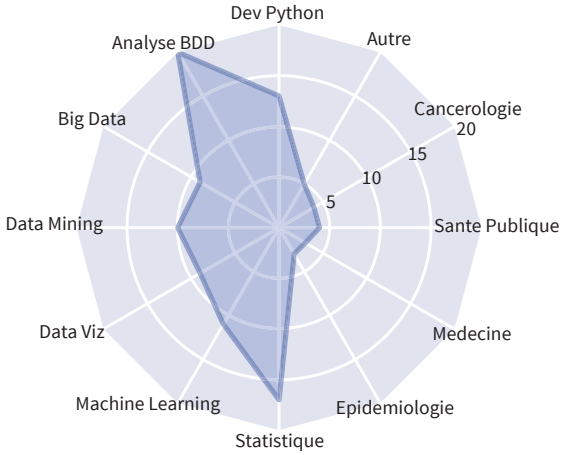
Figure n°10 : Distribution de l'activité des membres de l'équipe 5 au cours de la journée et de la semaine



Description : *Heatmap montrant le nombre d'edits par heure et jour de la semaine au sein de l'équipe 5.*

Source : wiki.epidemium.cc ; graphiques réalisés par les auteurs

Figure n°11 : Distribution des compétences au sein de l'équipe 5



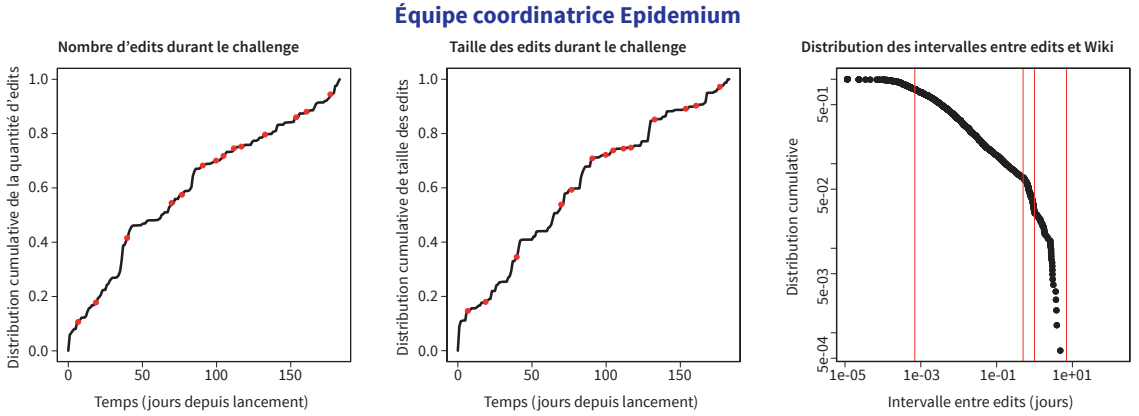
Description : *Représentation radar montrant la distribution des compétences individuelles au sein de l'équipe 5.*

Source : epidemium.cc, aptitudes déclarées par les participants lors de leur inscription ; graphiques réalisés par les auteurs

doute un cadre favorisant un travail organisé. Pour les autres équipes de la ***figure n°12*** *(voir ci-dessous)*, les équipes 11 et 12 expriment une certaine périodicité d'édition (« bosses » dans le panel de droite) alors que l'équipe 13 montre un travail par salves sans périodicité typique, similaire à l'équipe 5. Comme précédemment, un effet meetup apparaît, avec notamment une accélération après le point de mi-parcours, qui semble donc avoir porté ses fruits en tant qu'il avait été pensé comme un premier point d'étape pour les projets, venant ponctuer les six mois de Challenge.

Ainsi, la communauté a su faire émerger au sein des équipes-projets un travail collaboratif et productif dont l'équipe 5 est le symbole. Deux types de dynamiques émergent de ces études : une dynamique d'édition par salves qui est typique d'un travail

Figure n°12 : Comparaison de la dynamique temporelle du travail entre l'équipe Epidemium et les équipes n° 11, 12 et 13



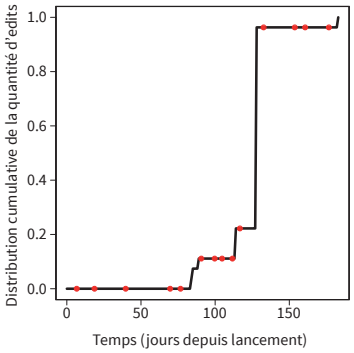
Description : *Diagrammes montrant la dynamique temporelle d'édition. Les diagrammes de gauche et du milieu montrent le nombre d'edits (resp. taille des edits) cumulés au cours du Challenge. Les points rouges correspondent aux événements Meetup. Le panel de droite correspond à la distribution cumulative de l'intervalle de temps entre deux edits (échelle log-log). Les lignes verticales rouges indiquent différentes échelles de temps : minute, demi-journée (12h), jour, semaine.*

Source : wiki.epidemium.cc , graphiques réalisés par les auteurs

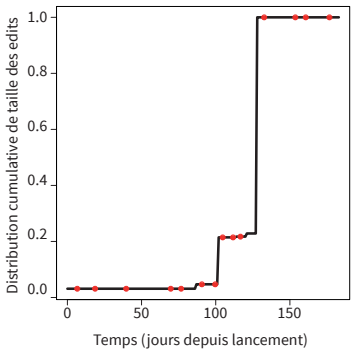
Suite page 49...

Équipe 11

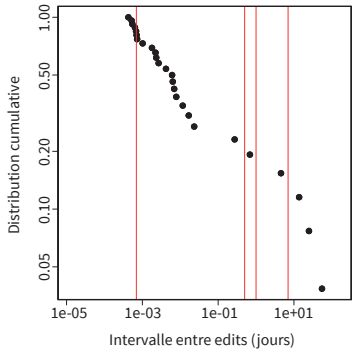
Nombre d'edits durant le challenge



Taille des edits durant le challenge

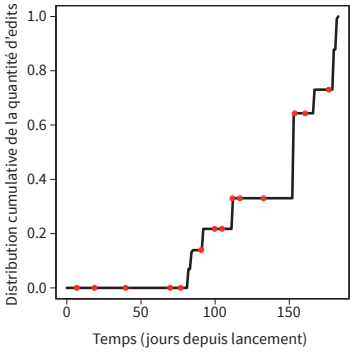


Distribution des intervalles entre edits et Wiki

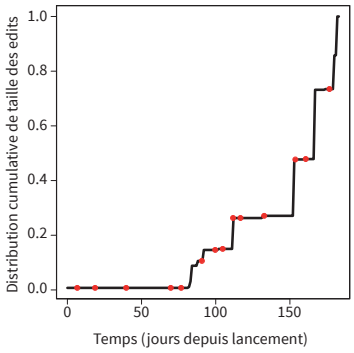


Équipe 12

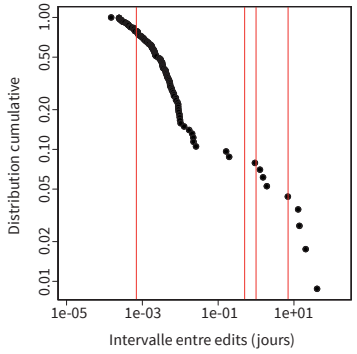
Nombre d'edits durant le challenge



Taille des edits durant le challenge

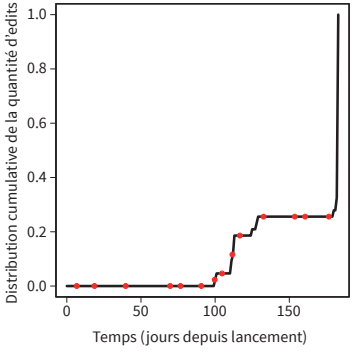


Distribution des intervalles entre edits et Wiki

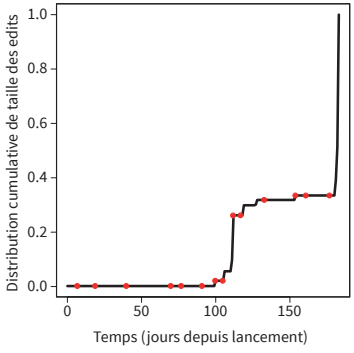


Équipe 13

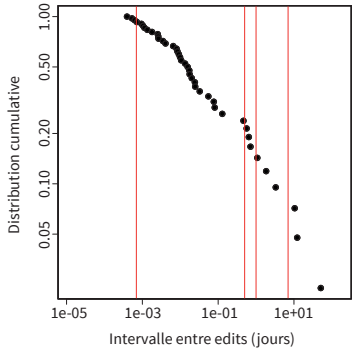
Nombre d'edits durant le challenge



Taille des edits durant le challenge



Distribution des intervalles entre edits et Wiki



continu sans synchronisation globale préalable (équipes 5 et 13) et une dynamique marquée par plus de périodicité, indiquant peut-être des rendez-vous pré-établis (équipes 11 et 12). Le rôle des événements organisés par Epidemium, et notamment le point de mi-parcours, est manifeste et a contribué à créer une dynamique d'engagement et de productivité.

Discussion

Les données recueillies par Epidemium nous ont permis de mener dans cet article une analyse de la dynamique d'auto-organisation des contributeurs à un projet *open science*. Nous avons pu dégager de cette analyse plusieurs réussites.

D'abord, nous avons pu observer l'effet positif d'une activation continue de la communauté sur la progression constante des inscriptions au Challenge4Cancer ainsi que de l'activité sur les différents outils en ligne mis à la disposition de la communauté. Cette corrélation démontre le grand potentiel de mobilisation de la thématique ainsi que l'efficacité des stratégies mises en place par l'équipe coordinatrice. L'engagement de cette communauté a été particulièrement fort lors de deux moments de synchronisation globale : le lancement du C4C et le point de mi-parcours. De plus, la communauté engagée s'est appropriée les différents outils que nous avons ici étudiés, proposés dans le cadre du C4C, à savoir le Wiki, le Q&A et le site web. L'analyse des données Wiki a montré une auto-organisation des acteurs en équipes aux méthodes de travail diverses. Une équipe de taille importante (plus de 30 contributeurs déclarés) a émergé, avec différents degrés d'implication, laissant apparaître une hiérarchie avec un petit groupe central. D'autres équipes de taille plus réduite ont montré une organisation plus simple avec une seule personne éditant la majorité du Wiki (données non montrées). Par ailleurs, deux dynamiques temporelles ont émergé, selon que les équipes ont travaillé de manière périodique ou de manière auto-organisée, sans temporalité pré-déterminée. Enfin, l'équipe coordinatrice d'Epidemium a réalisé un travail continu d'animation de la communauté tant dans la structuration du contenu du Wiki que dans l'animation

des événements. Le rôle de cette animation s'est fait ressentir dans la communauté et ses contributions : les événements ont cristallisé des échéances qui ont rythmé le travail des équipes et qui ont parfois permis une re-synchronisation des utilisateurs passifs en attente d'une possibilité d'engagement.

Ce Challenge constitue un événement sans précédent pour l'*open science*. Il fournit une preuve de concept à la lumière de laquelle il est possible de penser le futur d'un tel programme et les améliorations possibles. Nous recommandons tout particulièrement la mise en place d'un écosystème d'outils de travail connectés pour faciliter l'analyse en temps réel de la collaboration au sein des équipes. Cela serait bénéfique à la fois pour l'équipe coordinatrice et pour les participants. La première pourrait synchroniser son action aux besoins et à la dynamique de la communauté. Les seconds auraient une meilleure visibilité de l'ensemble des interactions en cours, favorisant ainsi leur engagement. La mise en place de ces outils permettrait une mise à l'échelle du programme et une nouvelle preuve de concept de l'animation d'une communauté massive et ouverte. ■

1. Outil interactif permettant de présenter un plan expérimental complet en mélangeant du code exécutable, de la documentation ainsi que des visualisations interactives.
2. Albert-László BARABASI, Network Science, Cambridge University Press, Cambridge (UK), 2016.

L'engagement de Roche

// **AUTEUR**

Stéphanie de Haldat

Filiale du cinquième investisseur mondial en R&D tous secteurs confondus, Roche France conjugue le modèle de recherche du Groupe en construisant des partenariats avec des acteurs de la santé de demain. À l'origine du projet Epidemium, deux convictions : pour trouver des solutions aux enjeux de santé de demain, il faut penser différemment, et les *open* big data sont une formidable opportunité pour la recherche épidémiologique.

La collaboration avec La Paillasse était une première pour nous. Pour un grand groupe Pharma, travailler avec un laboratoire ouvert et communautaire n'est a priori pas évident, mais cette première expérience s'est avérée être un succès et elle nous a appris beaucoup.

Le projet a créé un fort engouement chez les collaborateurs : un groupe projet dédié de 10 personnes, 24 ambassadeurs, plus de 20 collaborateurs impliqués, ... C'est en tout une cinquantaine d'employés Roche qui ont été sollicités et se sont investis dans

la démarche. Au moment du lancement d'Epidemium, nous avons senti un véritable enthousiasme pour le projet.

Néanmoins, nous avons été surpris de voir que proportionnellement peu de nos collaborateurs se sont véritablement mobilisés sur la plateforme collaborative. Plusieurs explications à ceci : pour les non spécialistes, la timidité face au sujet du big data et du *machine learning* qui peuvent être perçus comme étant difficiles à aborder et, bien sûr, la nécessité de libérer du temps dans des agendas chargés.

Au vu de cette expérience, nous envisageons les pistes suivantes pour une prochaine collaboration de ce type :

- acculturer nos collaborateurs non spécialistes au big data par une information pédagogique et ludique voire des sessions de formation ;
- se fixer pour objectif de constituer une équipe Roche dans Epidemium et de libérer du temps des collaborateurs pour qu'ils participent au projet.

D'un point de vue organisationnel, une personne a été nommée pour représenter Roche au sein de l'équipe projet Epidemium. Cette personne était l'interlocuteur principal de notre partenaire La Paillasse. Nous avons également créé une équipe projet dédiée à la Roche *open data base*, composante importante du projet Epidemium. Cette équipe était constituée d'une dizaine de personnes (médecin, juriste, biostatisticien, communicant, *etc*.) sous la responsabilité du chef de projet Epidemium.

Cette collaboration entre grand groupe Pharma et laboratoire ouvert et communautaire a fait apparaître des différences d'approche et de process. Les processus de travail que nous utilisons chez Roche sont très robustes et répondent aux exigences réglementaires auxquelles notre industrie est soumise. Ils sont également plus complexes et plus longs que ceux d'une équipe projet en mode start-up qui sont très agiles par essence. Ainsi le groupe projet Epidemium a pu référencer 21 000 jeux de données ouverts très rapidement. Parmi d'autres exemples, le groupe projet Epidemium a su faire réaliser très

« Epidemium fait converger l'inventivité du modèle de science ouverte prôné par La Paillasse et les objectifs en e-santé du leader mondial en cancérologie, Roche. »

Ewen Chardronnet
(Makery, 31/05/2016)

rapidement l'infographie du challenge en mobilisant des outils en *open source*. Nous pourrions développer ce type d'approche pour certaines prestations.

En termes de mode de collaboration et d'outils de travail, ce projet nous a appris à travailler d'une façon différente avec de nouveaux outils : Slack, Trello dont l'intérêt s'avère évident pour ce type de projet très décentralisé.

Notre collaboration avec La Paillasse a également montré des similitudes d'approche et de conviction. Nous nous sommes notamment retrouvés dans cette volonté commune de faire avancer la science au service du patient, dans la passion de la recherche et la rigueur scientifique.

Enfin, nous avons expérimenté à quel point la variété des profils et l'intelligence collective sont créatrices de valeur. Diversité et inclusion sont en effet des dimensions que Roche développe depuis plusieurs années pour aborder ce type de projet. ■

Les enseignements pour La Paillasse

// **AUTEUR**

Thomas Landrain

Epidemium a été pour La Paillasse une de ces opportunités qui vous changent la vie, si vous la saisissez. Pendant plus d'une année, nous avons travaillé d'arrache pied avec notre partenaire Roche pour montrer que la recherche scientifique, sur un sujet aussi complexe que l'épidémiologie du cancer, pouvait se faire de manière ouverte, coopérative et distribuée.

Voici les enseignements les plus importants qu'Epidemium nous a apporté :

- Les modèles ouverts représentent pour la recherche un intérêt majeur et ils sont de plus en plus incontournables pour les grands acteurs de l'innovation en santé, comme Roche a su le démontrer.
- L'importance de ce que nous appelons « la matière noire de la science », par analogie à ce qui compose la majorité

> *C'est une approche complètement nouvelle de l'épidémiologie, qui nécessite la collaboration de nombreux experts, dont des data scientists et des mathématiciens. Une expertise dont nous ne disposons pas en interne. »*
>
> ***Juliette Raynal***
> *(Industrie & Technologie, 08/07/2015)*

de l'univers et avec laquelle la matière traditionnelle n'interagit pas : ici, les données rarement partagées et le temps, l'expérience et les compétences de personnes hors du monde académique. Une telle approche inclusive permet de raccourcir les distances entres parties prenantes qui peuvent alors dialoguer sans intermédiation. Ceci n'aurait jamais été possible sans la fluidification des échanges et la transparence permises par notre approche *full open*, toute ressource offerte et produite au sein d'Epidemium ayant été ouverte et sans exclusivité.

- Grâce au contexte coopétitif de départ d'Epidemium, la plupart des participants se sont naturellement organisés en une coopération inter-équipe, et la majorité des contributions ont été orientées vers la construction de connaissances, de services et d'outils ayant un intérêt pour la communauté actuelle et future.

Ces points nous donnent à voir ce qu'une science sans opérateur pourrait être. ■

Les fiches

1a Le Comité d'éthique indépendant

Gilles Babinet	Entrepreneur, *Digital Champion France.*
Jérôme Béranger	Chercheur (PhD), Expert scientifique en Big Data, SI, Ethique et Réglementaire à KEOSYS.
Emmanuel Didier	Statisticien, Docteur en socio-économie de l'innovation et Professeur à l'ENSAE.
Muriel Londres	E-patiente, coordinatrice adjointe du collectif d'association de malades chroniques [im]Patients, Chroniques & Associés, Militante et bénévole dans l'association Vivre Sans Thyroïde.
Dr Cécile Monteil	Pédiatre urgentiste, Directrice médicale Ad Scientiam et Fondatrice de la communauté Eppocrate.
Pr Bernard Nordlinger	Service de Chirurgie Digestive et Oncologique à l'Hôpital Ambroise Paré et membre de l'Académie nationale de médecine.
Dr Jean-François Thébaut[1]	Cardiologue et membre du Collège de la Haute Autorité de Santé (HAS).
Me David Simhon	Avocat en droit de la santé et Président du Comité de Protection des Personnes Île-de-France III.
Pr Cédric Villani	Mathématicien, Professeur de l'Université de Lyon et directeur de l'Institut Henri Poincaré, médaille Fields 2010.

1. Jean-François Thébault était membre jusqu'au 31 janvier 2016 - démission pour raisons personnelles.

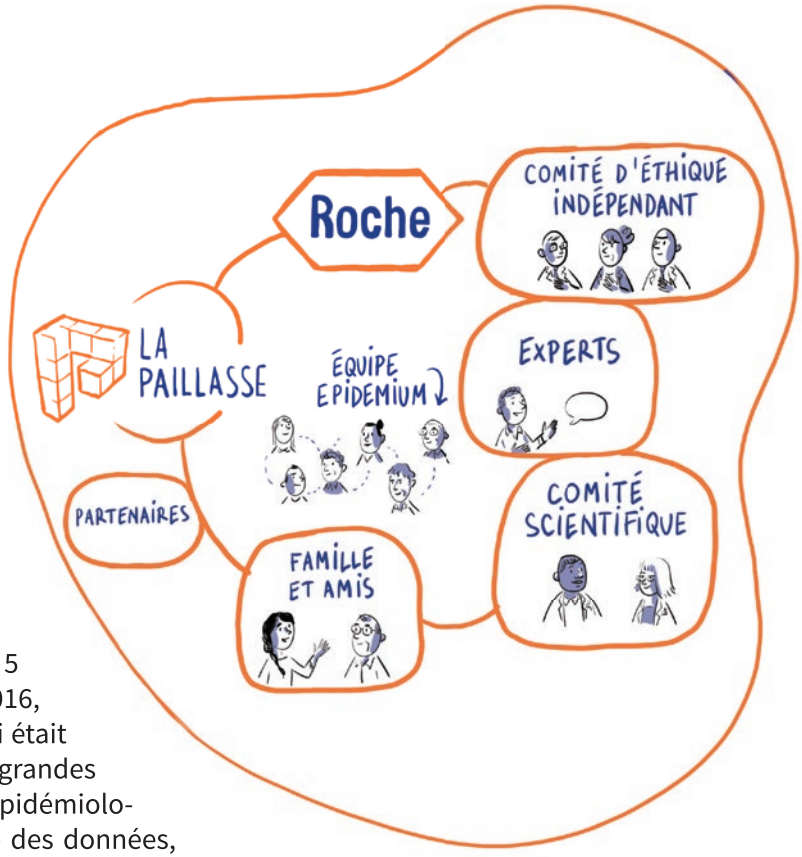
Le Comité scientifique

Aurélien Alvarez	Enseignant-chercheur en mathématiques, maître de conférences à l'Université d'Orléans, particulièrement intéressé par les systèmes dynamiques.
Dr Jean-Pierre Armand	Spécialiste en oncologie médicale, consultant senior à l'Institut Gustave Roussy et l'Institut Curie.
Djalel Benbouzid	Docteur en *machine learning*, post-doc au laboratoire LIP6, Université Pierre et Marie Curie.
Nicolas de Cordes	Vice-Président marketing anticipation du Groupe Orange, a notamment mis en place et géré les projets *Data4Development* en Côte-d'Ivoire et au Sénégal.
Dr Charles Ferté	Chef de clinique assistant en oncologie à l'Institut Gustave Roussy et expert en bio-informatique.
Pr Thomas Gauthier	Professeur de stratégie à la Haute école de gestion de Genève. Activités de recherche de son équipe : applications pratiques de la science des données et de la prospective à la prise de décision.
Dr Leila Kockler	Représentante Roche, directrice médicale projet à la Direction médicale de Roche France.
Thomas Landrain	Président & co-fondateur de La Paillasse.
Pr Philippe Ravaud	Professeur d'épidémiologie à l'Université Paris Descartes et à la Columbia University, directeur de recherche INSERM, directeur du Centre de Recherche en Épidémiologie et Statistique Sorbonne Paris Cité, directeur du centre d'épidémiologie clinique de l'Hôtel-Dieu (Paris), directeur de Cochrane Français, directeur du Centre EQUATOR France.

Epidemium dans toutes ses dimensions

// Epidemium :

Un programme de recherche scientifique collaboratif et ouvert à tous, initié par l'entreprise pharmaceutique Roche et le laboratoire communautaire La Paillasse. Son ambition est d'explorer le potentiel du big data en épidémiologie du cancer grâce à une communauté.



// Challenge4Cancer (C4C) :

Un grand appel à projets, du 5 novembre 2015 au 5 mai 2016, sous forme de coopétition, qui était structuré autour de quatre grandes thématiques, centrées sur l'épidémiologie du cancer et de la science des données, dans lesquelles des équipes pouvaient se constituer et développer un projet :

1. Comprendre la répartition du cancer dans le temps et dans l'espace.
2. Facteurs de risques et facteurs protecteurs du cancer.
3. Méta-épidémiologie : comprendre le cancer à partir de la littérature scientifique et médicale.
4. Changements environnementaux et cancer.

Ces thématiques étant complexes et la forme des projets n'étant pas présupposée, les équipes étaient encouragées à développer une certaine forme de transdisciplinarité en employant des compétences allant de la *data science* au domaine de la santé, en passant par les sciences sociales ou encore le design.

// Les Ressources :

Pour favoriser le succès du Challenge4Cancer et de ses participants, Epidemium a mis à leur disposition plusieurs ressources de type différent. D'abord, dans une dimension technique, ce sont plus de 21 000 jeux de données ouverts qui ont été assemblés et thématisés. Ces derniers ont été associés à un environnement d'analyse de données mis à disposition de la communauté par les partenaires techniques d'Epidemium : HyperCube, Dataiku et Teralab.

Epidemium a développé un volet événementiel, appelé Call4Debate, qui visait à stimuler la communauté, faire se rencontrer les participants et leur faire découvrir des solutions développées dans des thématiques connexes à Epidemium grâce à l'intervention des experts de l'écosystème du programme ou d'acteurs qui ont, dans leur bienveillance vis-à-vis du programme, accepté de le soutenir : Institut Curie, Paris-Saclay Center for Data Science, SchooLab, Bress, Quinten, Hacking Health, Global Knowledge, Club Jade, CapDigital, Cancer Campus, Wikimedia, ...

Enfin, pour présenter et favoriser la dynamique et le travail communautaire, plusieurs outils numériques ont été mis en place au fur et à mesure du programme et de l'expression des besoins des participants : un site web, un Wiki, un Q&A, un groupe et une page Facebook, un compte Twitter, un compte Meetup, ...

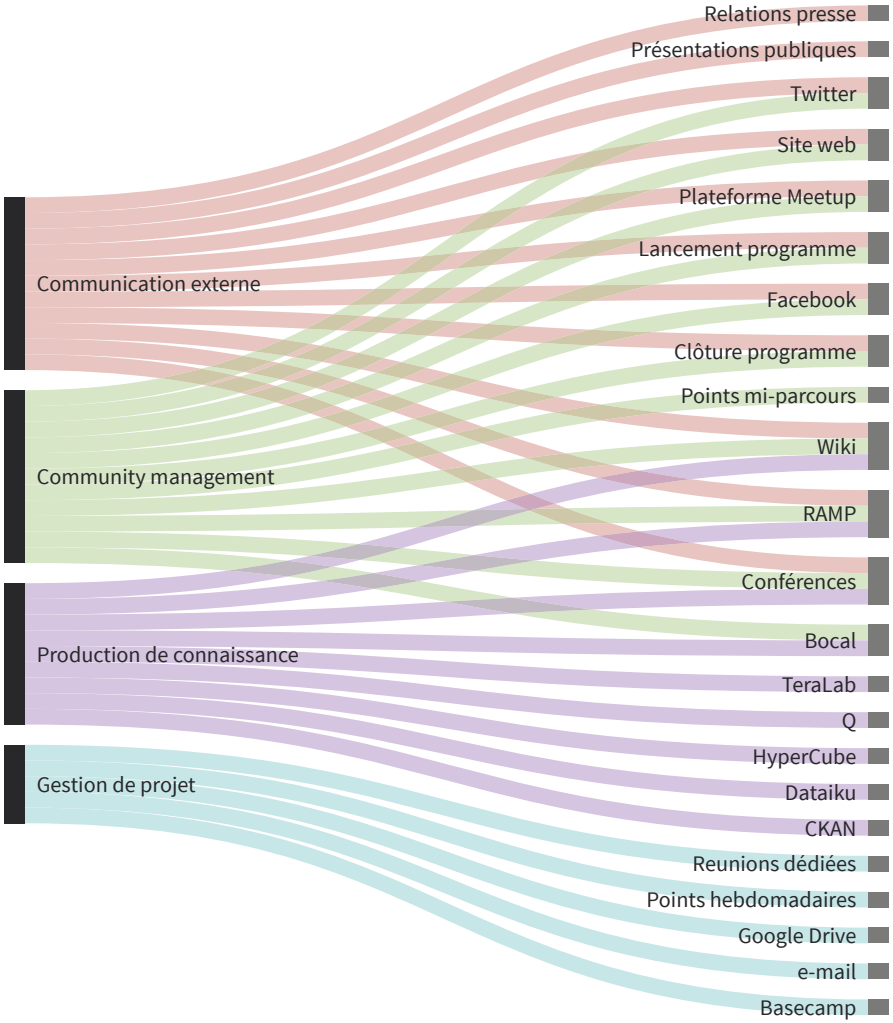
// Les Comités :

Pour encadrer et aider leur travail, un Comité d'éthique indépendant et un Comité scientifique, constitués en amont pour penser la constitution et le développement d'un tel programme, étaient présents. Ceux-ci, au terme des six mois de Challenge, se sont constitués en jury afin d'évaluer les projets soumis par les équipes.

Ainsi, pour résumer la première édition du Challenge4Cancer en quelques chiffres :

+ 20	678	15	8
événements	membres de la communauté	projets	projets finalistes

La boîte à outils d'Epidemium



Description : *Diagramme de Sankey représentant les outils mis en place par Epidemium et leurs fonctionnalités*

Source et réalisation : Équipe coordinatrice Epidemium.

Call4Debate 2015-2016

Date	Événement	Objet & Intervenant
06.10.15	**Conférence**	« Éthique et data de santé » **Jean-François Thébaut**, *Cardiologue, membre du collège de la HAS*
15.10.15	**Conférence**	« Le traitement des données de santé, enjeux et réalités » **Alexandre Templier**, *Spécialiste de la data science*
05.11.16	**Lancement du Challenge4Cancer**	Soirée destinée à la communauté Epidemium pour marquer l'ouverture du Challenge4Cancer, présenter les quatre thématiques et les modalités de participation
12.11.15	**Conférence**	« Les modèles d'épidémies sur base de données mobiles » **Nicolas de Cordes**, *Vice-président marketing anticipation du Groupe Orange* & **Stefania Rubrichi**, *Biomedical engineer & data scientist*
24.11.15	**Conférence**	« Prédire la survie des cancers du poumon de stade précoce » **Charles Ferté**, *Chef de clinique assistant en oncologie à l'IGR* & **Mathilde Bateson**, *Data scientist à l'Institut Hypercube*
10.12.15	**Bocal**	« Projets & constitution d'équipes »
12.12.15	**Conférence**	« *Open data* en cancérologie : un cas pratique » **Akpéli Nordor**, *Doctorant à l'Institut Curie et au Massachusetts General Hospital*
14.01.16	**Bocal**	« Développer, faire progresser, documenter vos projets »
21.01.16	**Conférence**	« *Open data* en santé : enjeux et débat » **Geoffrey Delcroix**, *Chargé d'études innovation et prospective à la CNIL* & **Jonathan Keller**, *Juriste de La Paillasse*
04.02.16	**Conférence**	« Que se cache-t-il derrière cette ère de l'*Open* (Big Data, Science, *etc.*) ? » **Guillaume Dumas**, *Co-fondateur de HackYourPhD* & **Célya Gruson-Daniel**, *Co-fondatrice de HackYourPhD*

	13.02.16	**RAMP**	Journée de travail coopétitif sur les données d'Epidemium
	18.02.16	**Conférence**	« Analyse des résultats du RAMP »
	25.02.16	**Bocal**	« Oncologie et épidémiologie »
	01.03.16	**Conférence**	« Un Google 3.0 du cancer : est-ce possible ? » **Alain Livartowski**, *Médecin à l'Institut Curie*
	12.03.16	**Point de mi-parcours**	Journée où les participants étaient invités à venir présenter leurs projets aux membres des comités et à échanger avec eux
	17.03.16	**Conférence**	« À la découverte de la *data science* + étude de cas en oncologie » **Amel Sahli**, *Docteur en Mathématiques et Market Manager chez Global Knowledge*
	07.04.16	**Conférence**	« Premières réalisations de la communauté Epidemium » **Méta-projet EpidemiumDB** & **Projet Viz4Cancer**
	14.04.16	**Conférence**	« Méthodes de travail collaboratif 3.0 : les intermédiaires pour faciliter la conception innovante » **Olga Kokshagina**, *Chercheuse à Mines ParisTech* & **Yohann Sitruk**, *Chercheur à Mines ParisTech*
	30.04.16	**RAMP**	Journée de travail coopétitif sur les données d'Epidemium & du méta-projet EpidemiumDB
	19.05.16	**Conférence**	« Meetup & Workshop in writing collaborative scientific articles - with Authorea » **Authorea**
	28.05.16	**Remise des prix & clôture du Challenge4Cancer**	Journée marquant officiellement la fin du Challenge4Cancer durant laquelle les équipes finalistes ont présenté publiquement et devant les comités leurs projets ; s'en est suivi la remise des prix et des mentions.

Type d'événements :

Conférence RAMP

Bocal Divers

Pour aller plus loin...

// Allier cancer et big data grâce à une méthodologie agile et flexible

- Olson G. (2000). “Distance Matters” dans Human-Computer Interaction, Volume 15, pp. 139–178, disponible en ligne <http://www.ics.uci.edu/~corps/phaseii/OlsonOlson-DistanceMatters-HCIJ.pdf>, dernière consultation le 30 novembre 2016.
- Lakhani K. R. (2016). “Managing Communities and Contests to Innovate with Crowds” dans Revolutionizing Innovation, pp. 109-134, Cambridge, MIT Press.
- Lakhani K. R. (2016). "Managing Communities and Contests to Innovate with Crowds” dans Revolutionizing Innovation: Users, Communities, and Open Innovation, Dietmar Harhoff D. et Lakhani K. R. (Dir), pp. 109–134. Cambridge, MIT Press.
- Houllier F. et Merilhou-Goudard JB. (2016). Les Sciences Participatives en France. Rapport février 2016, disponible en ligne <http://www.sciences-participatives.com/Rapport>, dernière consultation le 30 novembre 2016.

// Le Pouls du programme Epidemium

- Barabási A.L. (2016). Network Science. disponible en ligne <http://barabasi.com/networksciencebook/>, dernière consultation le 30 novembre 2016.
- Nielsen M. (2012). Reinventing discovery : the new era of networked science. Princeton University Press.
- Wuchty S., Jones B.F. et Uzzi B. (2007). “The increasing dominance of teams in production of knowledge” dans Science, mai 18; 316(5827):1036-9, disponible en ligne <http://www.kellogg.northwestern.edu/faculty/jones-ben/htm/Teams.ScienceExpress.pdf>, dernière consultation le 30 novembre 2016.
- Klug M. et Bagrow, J.P. (2016). “Understanding the group dynamics and success of teams” dans Royal Society Open Science, 6 avril 2016, disponible en ligne <http://rsos.royalsocietypublishing.org/content/3/4/160007>, dernière consultation le 30 novembre 2016.
- Börner K., Contractor N., Falk-Krzesinski H.J., Fiore S.M., Hall K.L., Keyton J., Spring B., Stokols D., Trochim W. et Uzzi B. (2010). “A multi-level systems perspective for the science of team science” dans Sci Transl Med, septembre 15, 2, 49cm24, disponible en ligne <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3527819/#>, dernière consultation le 30 novembre 2016.

#2

L'INNOVATION SCIENTIFIQUE ET MÉDICALE

Le progrès médical et l'amélioration des soins pour le patient sont le véritable centre névralgique d'Epidemium, si ce n'est sa raison d'être. Si l'on veut mieux soigner, guérir plus fréquemment et peut-être un jour vaincre le cancer, c'est par l'exploitation des quantités massives de données existantes, notamment épidémiologiques, qu'il faudra passer. La santé a en effet tout à gagner à s'emparer des nouvelles technologies et des nouvelles connaissances offertes par la science des données. À la clef ? Une médecine de précision et à la pointe de l'innovation !

// AUTEURS

Dr Charles Ferté / Pr Bernard Nordlinger / Dr Mehdi Benchoufi / Dr Perrine Créquit
Pr Philippe Ravaud / Muriel Londres / Dr Cécile Monteil / Équipe Baseline


DATA
DATA
CANCER

Quels usages de la science des données et du big data pour la santé ?

PATIENT EMPOWERMENT | GÉNOMIQUE | QUANTIFIED SELF | MACHINE LEARNING | MÉDECINE DE PRÉCISION

La santé est un enjeu majeur pour nos sociétés pour les années à venir. Contrairement à beaucoup d'autres industries, ce domaine n'a pas encore pris le virage du numérique, alors même que le potentiel offert par les nouvelles technologies, notamment pour transformer la prise en charge et la qualité des soins offerts aux patients, est particulièrement significatif. La numérisation de la santé génère quotidiennement de nouvelles données, qui permettront de faire entrer la médecine dans une nouvelle ère de soins plus personnalisés et plus justes.

// AUTEURS

Dr Charles Ferté | Pr Bernard Nordlinger

Bientôt, la santé de notre monde connecté ne sera plus celle du monde d'hier. Déjà, elle est le lieu privilégié où sont expérimentées les innovations les plus pointues dont nous entendons parler tous les jours. Objets connectés, intelligence artificielle, partage de données, big data, *blockchain*, … sont autant de termes qui pénètrent petit à petit le système de santé traditionnel pour le transformer.

Vers une prise de conscience

Des quantités massives de données de santé sont générées tous les jours, sur la prévalence et la mortalité des maladies, sur l'efficacité des traitements prescrits ou encore sur l'état de santé des citoyens et leur mode de vie. Une transformation s'opère néanmoins, donnant naissance à des interactions différentes et nouvelles entre le système de soins et le citoyen, patient ou non, qui répondent à des besoins concrets. Cela va des solutions de gestion d'agenda médical, comme Doctolib, ou de partage de dossiers médicaux, aux outils permettant d'analyser la voix et les mouvements fins dans le cadre de la maladie de Parkinson (projet mPower mené par Sage Bionetworks[1]), en passant par les *chatbots* offrant aux patients la possibilité de poser leurs questions à des "robots conversationnels" et de recevoir des réponses personnalisées, les APIs de traçabilité et de prévention d'épidémies ou encore les outils prédictifs proposant une médecine personnalisée, comme le fait MammaPrint[2] dans l'orientation et le choix des traitements pour les patientes souffrant d'un cancer du sein.

Ces nouveaux outils permettent aux hôpitaux de commencer à sortir du désert numérique qui les coupe des patients dès lors que ceux-ci ne sont plus physiquement présents dans les centres de soin. Cette discordance entre territoire physique et territoire numérique de l'hôpital est reconnue comme une limitation majeure de la qualité des soins et de l'efficacité du suivi des patients. Il faut donc que l'hôpital d'aujourd'hui devienne un acteur hybride, à la fois acteur numérique et de la vie réelle, pour les citoyens. Ainsi, une fois que les professionnels de santé (centres de soins, hôpitaux et soignants) parviendront à offrir aux citoyens la possibilité de rester connectés en dehors

même des lieux de soin, la solution d'une interaction en continu avec le patient pourra être mise en place, offrant au système de santé des solutions aux enjeux clefs d'aujourd'hui comme le *reporting* des effets secondaires, l'éducation et la prévention, l'information sur les maladies, *etc.*

Au-delà du bénéfice évident acquis grâce à une meilleure interaction patients-médecins pour la qualité des soins et le type de services apportés, ces outils permettent également au monde médical et à la société de collecter de très nombreuses données épidémiologiques et environnementales. Ces dernières sont le terreau nécessaire aux évolutions majeures du système de santé traditionnel, que nous allons présenter. Elles engendrent aussi l'apparition de nombreux défis complexes, dans l'objectif de tendre toujours plus vers l'idéal d'une médecine personnalisée, adaptée à chaque patient et appréhendant chaque maladie comme un cas unique.

La révolution, ce n'est pas seulement le diagnostic : c'est le traitement, le suivi, le pronostic et la prévention

La première évolution majeure des données en santé est l'explosion du nombre de données disponibles. La génomique représente une vaste source d'informations pour les médecins et les chercheurs dont ils se sont largement emparés depuis maintenant plusieurs années.

Qui n'a pas entendu parler du *Human Genome Project*, l'un des, si ce n'est le plus important, événements scientifiques de la génération actuelle ? En cancérologie plus spécifiquement, *The Cancer Genome Atlas* (TCGA)[3] et l'*International Cancer Genome Consortium* (ICGC)[4] sont deux grands programmes de séquençage qui ont généré de considérables quantités de données publiques grâce à des cohortes de plusieurs centaines de patients. TCGA, par exemple, a généré une carte multi-dimensionnelle des mutations génétiques pour trente-trois types tumoraux avec séquençage ADN, ARN, RPPA, *etc.* Ces programmes internationaux ont pour ambition de mieux comprendre le cancer grâce au séquençage du génome

complet de dizaines de tumeurs différentes. La communauté scientifique peut alors s'emparer des immenses bases de données ainsi mises à sa disposition pour accélérer la recherche contre le cancer.

L'ère de la génomique représente l'avenir immédiat et à moyen terme mais, parallèlement, d'autres types de données connaissent une importance croissante. Ainsi, le *free text*, soit l'ensemble des écrits produits par les professionnels de santé, représente une nouvelle source majeure de données permettant de démultiplier les informations disponibles sur la maladie et le traitement. C'est le cas également des données d'imagerie, qui sont de plus en plus nombreuses, grâce notamment à l'amélioration des méthodes d'imagerie médicale. En 2015 aux États-Unis, 80 millions de CT-scan ont été effectués, quand ce chiffre était quatre fois inférieur vingt ans plus tôt .

On assiste également à une augmentation considérable des données dites *quantified self*, c'est-à-dire collectées et fournies par le citoyen lui-même, parfois à son insu d'ailleurs, par ce que l'on nomme les *weareables*, ces « objets connectés portables » (le terme n'a pas encore d'équivalent en français) et les applications qui leur sont liées.

Or, lorsque de telles quantités de données sont disponibles, avant de les analyser, l'enjeu est d'abord de les rendre accessibles en mettant à disposition de tous ces jeux de données collectés par les organismes de santé et de recherche, par les institutions publiques et par les entreprises privées détenant les objets connectés, les applications mobiles et autres *weareables*. Pourquoi le partage des données générées est-il essentiel ? Parce qu'il entraîne une démultiplication de leur impact en permettant à chacun de s'en emparer, de les préprocesser, de les analyser puis de les interpréter. Cette analyse est justement rendue possible aujourd'hui grâce à la création de nouveaux algorithmes de *machine learning* plus efficaces, plus précis et plus démocratiques. Des projets majeurs comme Watson d'IBM [5] et *Deepmind* [6] de Google, après avoir fait leurs preuves dans d'autres domaines (aux échecs ou au jeu de Go par exemple), se positionnent tous deux sur le domaine de la santé pour répondre aux immenses quantités de

> *Certaines de ces applications se concrétiseront et d'autres non, soit pour des raisons techniques soit parce que tout progrès est plus hasardeux et difficile quand il concerne l'être humain que le commerce ou le transport en taxi. »*
>
> ***Pr Bernard Nordlinger***

Quels usages de la science des données et du big data pour la santé ?

données en circulation. Comme si les champs d'application sur lesquelles ont été exercées ces super-intelligences avaient servi d'entraînement à leur puissance de calcul avant de s'attaquer à des enjeux plus sérieux tels que la santé publique. Le cancer est donc le nouvel ennemi commun de ces intelligences artificielles et de très puissants moyens sont concentrés sur cet objectif par de grands groupes qui se donnent pour ambition de mieux répondre aux besoins des citoyens. En août 2016, Watson a ainsi diagnostiqué un cas de leucémie qui n'avait pas été détecté par l'intelligence humaine[7], prouvant que l'intelligence artificielle mise au service de la santé sera source d'amélioration considérable des soins pour les patients et de gain sans précédent en santé publique. L'Université de Tokyo a, quant à elle, indiqué en septembre 2016 que Watson avait aidé à diagnostiquer et à traiter les patients atteints de cancer dans 80% des cas proposés à son analyse[8].

Parallèlement, l'accès au *cloud* et le développement de nouveaux outils ouvrent la possibilité pour chacun de faire, soi-même, ce qui était jusqu'ici réservé aux entreprises détenant les technologies. La diminution du prix et l'augmentation de l'efficience des technologies permettent une véritable démocratisation de leurs usages. De nombreux hébergeurs voient le jour sur le *cloud* en complément des majors que sont Amazon, Azure et Google, avec notamment des *clouds* hybrides proposant des services plus personnalisés que chacun peut s'approprier. Toutefois, il convient alors d'être très attentifs à la sécurité des données car beaucoup sont identifiantes, c'est-à-dire qu'en les recoupant, il est possible de retrouver l'identité du patient auprès duquel elles ont été collectées. C'est pourquoi, aujourd'hui, l'hébergement des données est réservé aux acteurs capables de respecter l'article L.1111-8 du Code de la santé publique.

Le corollaire de ce partage est que l'on voit se développer de nombreuses initiatives collaboratives qui profitent de ces données ouvertes. Sous le format, le plus souvent, de *data challenges*, des communautés se fédèrent en ligne et hors ligne, ouvertes à tous, autour d'un objectif commun et dans un cadre partagé. C'est le cas de groupes ou plateformes comme Synapse[9], Kaggle[10] ou ici d'Epidemium, où s'opèrent partage

L'article L.1111-8 du Code de la santé publique précise les conditions dans lesquelles les données de santé peuvent être confiées à un hébergeur.

- *La personne concernée par les données de santé doit avoir consenti expressément à l'hébergement de ses données.*
- *L'hébergeur doit être agréé pour son activité.*
- *L'hébergeur est soumis aux règles de confidentialité prévues à l'article L.1110-4 du Code de la santé publique et à des référentiels d'interopérabilité et de sécurité.*
- *Lorsque les professionnels de santé ou les établissements de santé hébergent leurs propres données de santé, ils ne sont pas soumis à l'agrément et ne sont pas tenus de recueillir le consentement exprès de l'intéressé pour conserver ces données*

de données, mise en ligne des outils nécessaires, co-création d'algorithmes, et surtout mise en commun des savoir-faire et des compétences. En développant un travail collaboratif organisé autour de l'épidémiologie du cancer, le programme Epidemium et son Challenge4Cancer ont montré combien la recherche scientifique et les patients pouvaient bénéficier de plus d'interdisciplinarité et d'ouverture. C'est ce genre de programmes collaboratifs qui peuvent changer la donne pour le milieu de la recherche traditionnelle, en lui apportant ce dont il manque aujourd'hui : des expertises plus variées, puisqu'un graphiste y aura autant sa place qu'un *data scientist*, la mise en commun des savoir-faire sans enjeu de pouvoir et le partage sans restriction des résultats avec le reste de la communauté.

Vers une médecine personnalisée, un parcours semé de défis

Pour autant aujourd'hui, aucun outil, aucune plateforme, n'a encore changé la pratique de tous les jours pour le monde de la santé. Comment alors ouvrir la voie à une véritable médecine personnalisée ? Est-ce utopique d'imaginer une médecine où

> *« À mon sens le plus grand challenge sera de pouvoir croiser les informations cliniques ou épidémiologiques, c'est-à-dire ce qui est exprimé, le phénotype, avec les données génétiques. On n'y est pas encore. On connaît de plus en plus de biomarqueurs (sur un nombre limité de gènes) qui permettent d'adapter certains traitements anticancéreux à ceux qui ont une chance de pouvoir en profiter. Le séquençage du génome est devenu une pratique courante mais le diagnostic moléculaire n'a que peu d'applications actuelles dans le traitement des cancers. »*
>
> ***Pr Bernard Nordlinger***

Quels usages de la science des données et du big data pour la santé ?

chacun aurait accès à des outils permettant de choisir son traitement en fonction des meilleures prédictions faites grâce au big data, prenant en compte un ensemble de critères variés allant de la génétique au mode d'alimentation ? Ou encore de dépister de manière précoce la maladie d'un individu grâce aux requêtes qu'il a effectuées sur les moteurs de recherche en ligne ? Ces deux cas particuliers existent déjà mais restent isolés. Pour les démocratiser, de nombreux défis sont à relever car, si les programmes de traitement personnalisé font preuve de résultats concluants, laissant espérer de belles promesses, ils restent encore confidentiels et les quelques succès individuels obtenus sont loin d'être une généralité pour les patients.

Les défis pour parvenir à des résultats positifs en médecine personnalisée sont d'abord techniques : comment tirer des enseignements utiles de données très dispersées ? Les données collectées sont hétérogènes par leur nature (génomiques, physiologiques, biologiques, sociales, environnementales…), leur format (texte, valeurs numériques, signaux, images 2D et 3D, séquences génomiques…), leur dispersion au sein de plusieurs systèmes d'information (groupes hospitaliers, laboratoires de recherche, bases publiques, sociétés privées…). Or, en big data, il n'y a pas d'intelligence sans apprentissage. La grande fragmentation des données demande donc d'inventer des systèmes complexes afin de réussir l'intégration de données de nature et de source différentes. Dans la même logique, on voit également se développer un besoin croissant en algorithmes et en capacité de stockage et de calcul de ces bases de données.

Sur le plan technique également, l'un des problèmes soulevés est que les structures de santé, volontaires face à l'arrivée de telles innovations, ont toutes développé leur propre système de santé indépendamment de celui des autres structures. Les systèmes ainsi créés ne sont donc pas interopérables, ce qui représente un frein important au partage des données. Comment s'assurer en effet, lorsqu'un patient effectue des consultations auprès de structures différentes, que ses données sont effectivement transmises sans perte ni problème de format, utilisables par des professionnels de santé n'ayant pas forcément le temps ni la possibilité d'échanger entre eux ?

Rendre le citoyen maître de ses données et acteur de leur partage est également un défi majeur. C'est ce qu'on appelle le *patient empowerment*. Dans un objectif de santé publique, chacun devrait être conscient de la richesse des données qu'il crée au quotidien et de leur utilité potentielle pour le monde médical, la compréhension et donc le traitement de maladies complexes comme le cancer. Idéalement, demain, les programmes de recherche n'auront plus besoin de créer des cohortes de citoyens volontaires pour donner leurs données, mais pourront puiser dans la richesse des données pré-existantes grâce à la sensibilisation des citoyens au quotidien.

Pour réussir à disposer de données suffisantes en nombre et en diversité, les institutions doivent alors s'emparer du sujet de l'éthique. En effet, l'utilisation des données issues des essais cliniques ou collectées lors du parcours de soin entraîne une obligation d'information pour le patient : il est nécessaire

Quels usages de la science des données et du big data pour la santé ?

d'expliquer quelles données seront collectées, comment elles seront anonymisées, dans quel cadre elles seront stockées et pour quel objectif elles seront amenées à être utilisées. Le travail de clarification et d'éducation est essentiel afin que le grand public comprenne l'intérêt que représentent leurs données pour la recherche et la santé publique. Ces principes éthiques de transparence dans le recueil, l'analyse et le traitement des données nécessitent donc une surcharge de travail et de processus parfois lourds à supporter pour le système de santé.

Collecter des masses de données en donnant confiance aux citoyens pour les partager représente enfin un défi de sécurité. Pas forcément, comme il est habituel de le penser, celui du piratage des données, mais plutôt celui de leur non-corruption. En 1996, les États-Unis ont instauré le *Health Insurance Portability & Accountability Act* (HIPAA)[11] qui impose la mise en œuvre de mesures de sécurité et de respect de la vie privée pour la création, la conservation et la transmission des données de santé personnelles. De son côté, l'Union européenne a adopté en 1995 la Directive sur la protection des données, créant un environnement homogène dans toute l'UE. Cependant, si la collecte et le traitement des données personnelles sont bien soumis à des obligations, une loi ne peut empêcher une faille technologique ou des actes malveillants. Une solution innovante intéressante apparaît alors avec la promesse du *blockchain* qui pourrait être mis en place dans tout le système de santé[12]. Toutefois, cette technologie est-elle mature ? Et surtout, le système de santé est-il prêt à opérer une telle innovation dans son mode de fonctionnement ? Il faudra probablement plusieurs années avant que l'on assiste à une généralisation du *blockchain* dans le domaine la santé mais cela n'en reste pas moins un bel objectif à poursuivre.

À l'heure de l'ubérisation croissante de très nombreux domaines économiques, on tarde toujours à voir apparaître les licornes en médecine, ces start-ups des nouvelles technologies, très innovantes, à la croissance rapide et qui ont atteint un milliard de dollars de valorisation. Le milieu médical nécessite, en effet, de conserver un compromis important entre les innovations et leurs usages, qui limite l'implémentation hasardeuse des nouvelles technologies. Bien entendu, le big data est là

pour accroître l'efficience des métiers, des traitements et de la prévention, que ce soit par une meilleure prédiction, la diminution des coûts par la limitation des actes ou examens inutiles, *etc.*, ou par une meilleure connaissance des maladies.

Toutefois, l'usage des technologies du big data ne peut pas se faire au prix d'un moindre niveau de preuve, ce qui signifierait un moindre niveau de validation des résultats scientifiques, ou d'une moindre sécurité des données. C'est pour cette raison que l'implémentation du big data en santé prend du temps. Nombreux sont ceux qui s'alarment pourtant aujourd'hui car les aspects concrets des promesses de médecine de précision tardent à venir. Cependant, le mouvement est bel et bien lancé et l'on ne reviendra pas en arrière. C'est un mouvement de fond dont l'*empowerment* des patients en particulier sera une composante très forte, poussant le monde médical à des performances meilleures, plus rationnelles et plus ouvertes. Nous n'en sommes qu'au début de l'histoire. ■

1. Site mPower mobile Parkinson Disease Study <http://parkinsonmpower.org/>, dernière consultation le 30 novembre 2016.
2. Site Agendia, MAMMAPRINT® 70-GENE BREAST CANCER RECURRENCE ASSAY <www.agendia.com>, dernière consultation le 30 novembre 2016.
3. Site The Cancer Genome Atlas <http://cancergenome.nih.gov>, dernière consultation le 30 novembre 2016.
4. Site International Cancer Genome Consortium <http://icgc.org>, dernière consultation le 30 novembre 2016.
5. Article "Un vaste champ d'applications dans la vie de tous les jours" sur le site d'IBM <http://www-05.ibm.com/fr/watson/>, dernière consultation le 30 novembre 2016.
6. DeepMind est une entreprise britannique spécialisée dans l'intelligence artificielle rachetée en 2014 par Google <https://fr.wikipedia.org/wiki/Google_DeepMind>, dernière consultation le 30 novembre 2016.
7. Humanoides, IBM Watson diagnostique avec sucès un cas de leucémie au Japon, mise en ligne le 10/08/2016 <https://humanoides.fr/ibm-watson-japon-leucemie/>, dernière consultation le 30 novembre 2016.
8. T. KAWAMURA, "Big data system shows promise in helping cancer patients at Todai", The Asahi Shimbun, 19 septembre 2016, disponible en ligne <http://www.asahi.com/ajw/articles/AJ201609190064.html>, dernière consultation le 30 novembre 2016.
9. SAGE SYNAPSE, <https://www.synapse.org>, dernière consultation le 30 novembre 2016.
10. KAGGLE, Your home for data science <https://www.kaggle.com>, dernière consultation le 30 novembre 2016.
11. European Union Agency for Network and Information Security (ENISA), Health Insurance Portability and Accountability Act, disponible sur <www.enisa.europa.eu>, dernière consultation le 30 novembre 2016.
12. Lire aussi D. SCHUYLER, Is the Blockchain a Potential Cure for Securing Health care data? <http://leavittpartners.com/2016/09/is-the-blockchain-a-potential-cure-for-securing-health-care-data/>, dernière consultation le 30 novembre 2016.

« Le problème concerne le recueil des données personnelles, consenti ou non, quand on sait les difficultés d'une véritable anonymisation, mais aussi l'usage de ces données et pas seulement pour les compagnies d'assurance, pour prendre l'exemple le plus courant. »

Pr Bernard Nordlinger

Crowdsourcer une épidémiologie du cancer

EPIDÉMIOLOGIE CROWDSOURCING CROWDACTING

MICRO-TASKING MEGA-TASKING

La santé, et singulièrement l'épidémiologie, sont touchées en profondeur par les mutations technologiques en cours, et ce, sous le double impact de la production d'un volume inégalé de données et de l'implication croissante de communautés. Faut-il encore trouver les voies et méthodes pour en exprimer le potentiel. De nombreuses techniques d'engagement des communautés sont aujourd'hui pratiquées (micro-tasking, mega-tasking) et laissent augurer d'approches épidémiologiques originales, larges, distribuées, empruntes de dynamiques réticulaires et sociales, augmentées par le temps réel et, tel que le projet Epidemium, en défriche la promesse, puissamment inclusives.

// **AUTEURS**

Dr Mehdi Benchoufi / Dr Perrine Créquit / Pr Philippe Ravaud

Le monde de la santé est bouleversé dans ses pratiques et secoué dans ses usages par le double impact des masses de données et des masses d'individus qui s'invitent dans le jeu de la recherche bio-médicale, c'est-à-dire par l'émergence du big data et par l'implication croissante et communautaire de la société civile, lesquels se hissent à la hauteur des défis de la médecine contemporaine. L'épidémiologie, par essence tournée sur l'usage de données et la capture d'initiatives extérieures à celles menées par son corps de praticiens usuels, est aussi bien un poste d'observation qu'un champ d'expérimentation de cette nouvelle donne. L'épidémiologie vit un changement que nous pouvons, sans trop nous avancer, qualifier de paradigmatique[1].

En effet, les exemples de nouveaux usages en épidémiologie, fruits d'approches singulières de la part d'acteurs tiers au système de santé, sont nombreux. D'ailleurs, ils sont souvent présentés comme des exemples significatifs du potentiel né de la fertilisation croisée de données gigantesques, de la capacité de calcul nécessaire à leur traitement et de la masse d'internautes capables de s'impliquer dans leur analyse.

Lorsque le potentiel des données massives est catalysé par des communautés d'individus aptes à les travailler, le fruit de cette composition, soit le *crowdsourcing*[2], peut doter l'épidémiologie de moyens nouveaux et intéressants. Le *crowdsourcing* est efficace en ce qu'il permet la mobilisation et la mutualisation d'une force de travail largement distribuée. Sa forme minimale mais la plus répandue, dite *micro-tasking*, qui consiste à subdiviser une tâche complexe en une somme de tâches élémentaires, est un des modes de recours au *crowdsourcing* le plus fréquent en recherche biomédicale. Il se fait selon un jeu de contre-parties, le plus souvent financières.

Il est aussi une autre forme de *crowdsourcing*, parfois qualifiée de *mega-tasking*, témoignant de la volonté de s'impliquer, de se rendre utile en contribuant, de mettre à profit des compétences non médicales à la faveur d'un enjeu de société. C'est une marque de notre époque. Epidemium en aura été un exemple concluant. L'envie d'engagement, l'idée de se hisser sans complexe à la hauteur de challenges ambitieux, que nous

> *« Les big data vont permettre de déceler les facteurs responsables de l'émergence du cancer chez les patients, et de modifier notre approche de la santé publique en France. »*
>
> ***Muriel Londres***
> *Membre du Comité d'éthique indépendant*

avons pu apprécier tout au long du Challenge4Cancer, est une manifestation d'un mouvement général, parfois appelé *Do It Yourself* (DIY), témoignant de l'idée que si les problèmes nous concernent tous, alors les solutions appartiennent à chacun.

Du crowdsourcing au crowdacting

Internet est *crowdsourcing*. Avec la somme de nos interactions, nous l'alimentons quotidiennement d'une masse de données considérables qui sont qualifiées par de grandes plateformes, pressées, raffinées, transformées pour en extraire la valeur d'usage, c'est-à-dire la valeur capitalistique, et pas assez encore la valeur scientifique. L'abondance de ces petabytes de données peut être vu comme un *crowdsourcing* naturel ou passif. En effet, à l'ère d'Internet, tout est « donnée », et d'ailleurs tout est donné par les internautes. Nous distinguerons donc un *crowdsourcing* qui est la substance active d'Internet, d'un *crowdsourcing* plus volontaire que nous qualifierons de *crowdacting*.

L'épidémiologie est, entre autres, l'étude des déterminants des maladies. Elle ne restera pas insensible à l'impact des évolutions que nous évoquons. Ses moyens de connaissance et ses moyens d'interventions, dans le temps et dans l'espace, sont aujourd'hui augmentés. D'une part, les données élargissent et approfondissent nos clés de compréhension de la genèse des maladies. D'autre part, la dématérialisation des supports dont la transition numérique est le média, la reformulation de l'espace, aujourd'hui sans territoire, et du temps, présent perpétuel, donnent à notre discipline une capacité d'action dite en temps réel, dont on apprécie le potentiel dès lors que l'on imagine le contrôle de la propagation des maladies.

Au-delà, on comprend que, d'une part, la diversité et la masse des données produites par nos systèmes confèrent les outils pour amplifier considérablement la connaissance de notre environnement, de nos comportements, et que, d'autre part, la mutualisation des efforts de recherche, impliquant aussi bien les circuits académiques que des citoyens experts, est une opportunité décisive. Ces citoyens avides et acteurs

d'une science plus ouverte sont à la fois force d'appui et de démultiplication du travail, dans une optique de délégation de micro-tâches, et peuvent aussi bien, selon le modèle de type challenge dont Epidemium est un avatar, fournir des moyens de connaissances hétérodoxes.

Exemples de crowdsourcing

Crowdsourcing passif

// Quelques exemples

Par des moyens de *crowdsourcing*, l'initiative HealthMap[3] a détecté une fièvre suspecte en Afrique avant même que les autorités sanitaires ne soient alertées par ce que l'on découvrira être la fièvre Ebola. La méthode consiste ici à opérer une analyse continue d'une masse de données hétérogènes collectées depuis des sources d'informations variées : sites experts, blogs, réseaux sociaux, forum de santé. Ces dernières sources étant typiquement le fruit d'un *crowdsourcing* passif ou plutôt d'un *crowdsourcing* au sens littéral : rassembler les données de la foule depuis la source dans lesquelles elles sont produites.

Indiquons que les exemples les plus inspirants ne procèdent pas nécessairement de succès fracassants mais éclairent des approches expérimentales prometteuses, pavent des chemins nouveaux et sont parfois les échecs qui annoncent les victoires. Parmi ceux-là figure le très abondamment commenté Google Flu[4]. Au terme de cette expérience, Google Flu n'a pas réussi, comme cela en était le propos, à anticiper ni prédire la diffusion de la grippe. En revanche, nous tiendrons pour intéressante l'idée de pouvoir se donner une intuition du phénomène en contournant le labeur de la récolte « manuelle » et l'agrégation des données, de la synthèse minutieuse de ces informations, de leur analyse par des cellules de veille expertes. C'est là l'économie de moyens qui permet de compléter le travail des hommes par une approche automatisée via des algorithmes. Indiquons tout de même que, pendant deux années consécutives, Google Flu a réussi à prédire fidèlement, en avance sur les systèmes de veille sanitaire, l'évolution de la grippe. Disons

> *En effet, dans le sport collectif et décentralisé que deviendrait la recherche en épidémiologie, l'épidémiologiste est aussi community manager. »*

que l'algorithme manque la cible mais ce que nous en retiendrons, c'est qu'il n'en est pas loin.

// Un crowdsourcing augmenté : le potentiel du machine learning

À ce stade, nous devons faire état d'une approche technologique essentielle qu'est le *machine learning*, lequel consiste à éduquer des ordinateurs en leur faisant apprendre des données, qui en retour acquièrent de l'expérience et affûtent leur capacité d'analyse préventive, prédictive, voire pour certains cognitive. Ces techniques sont particulièrement requérantes et dépendantes du volume de données. Les masses de données de santé dont disposent les systèmes sanitaires peuvent être perçues comme le fruit d'un *crowdsourcing* passif et sont alors un élément précieux qui offre des perspectives d'innovation thérapeutique à pharmacopée constante.

Ces techniques valent des résultats spectaculaires : la victoire de l'intelligence artificielle de Google dans le jeu de Go, la détection automatisée de tumeurs à partir d'images scannographiques, ...

Bien sûr, le volume des données ne confère pas en soi un avantage d'analyse statistique. En revanche, certains algorithmes tirent leur puissance au fur et à mesure qu'ils sont dotés en données. Il en va ainsi d'un domaine particulièrement populaire, à savoir les réseaux de neurones, dont certains résultats sont tout à fait spectaculaires. Cette approche est d'une certaine façon bio-mimétique, elle rassemble des neurones comme autant d'unités de calculs dont les règles de calcul sont précises et les paramètres fluctuent au fur et à mesure des données processées.

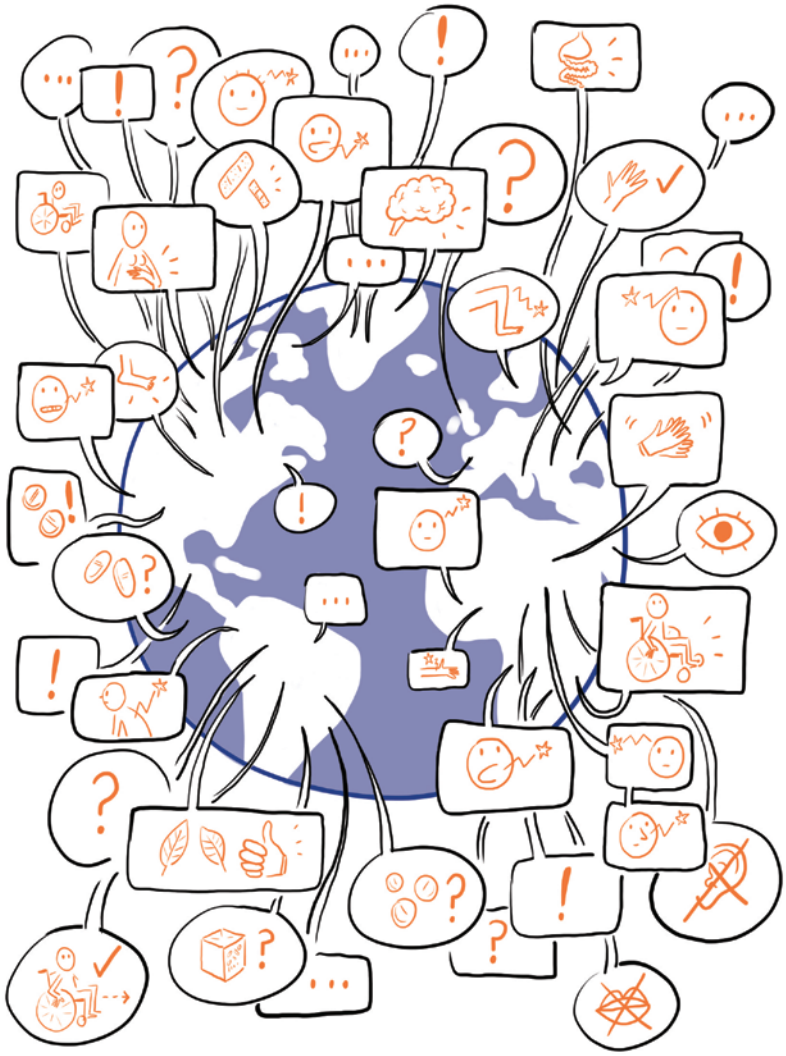
Crowdsourcing actif

// Micro-tasking

Le *micro-tasking* illustre bien l'engagement des communautés à participer à la co-construction de leur santé, ainsi que les effets de levier importants que représentent ces mobilisations pour les chercheurs, permettant d'accéder à la réalisation de tâches jusqu'alors difficiles, moins par leur complexité intrinsèque

que par le temps-homme considérable qu'elles requièrent. Le projet Embase, conçu par la Collaboration Cochrane, est un bon exemple de division et de parallélisation de tâches, consistant à identifier les rapports d'essais cliniques provenant d'Embase, une vaste base de données bibliographiques, à des fins de publication dans le registre central de la Cochrane. Il est notable qu'aucune expertise préalable n'est attendue des participants, lesquels peuvent être naïfs de toute connaissance médicale.

Dans l'esprit d'une recherche médicale plus inclusive, les projets Compare ou Nutrinet permettent à des patients de s'engager dans des études cliniques en ligne et de participer ainsi à la marche d'une recherche médicale plus ouverte, esquissant la promesse d'une médecine à la fois participative et contributive. En effet, dans Compare, les patients atteints de maladies chroniques partagent leur données, répondent à des questionnaires, participent au choix des sujets de recherche, l'objectif étant de mobiliser jusqu'à 200 000 patients.



Selon une approche originale, Epidemium montre qu'autour d'un Challenge, les communautés trouvent des formes d'organisation sinon spontanées, au moins proches de celles

auxquelles les conforment le plus les compétences qu'elles rassemblent. Ainsi, certaines équipes, à défaut d'expertise médicale ou informatique, se sont lancées dans un travail de collection et de nettoyage des données remarquable. Par exemple, le projet Baseline a pu développer une riche base de données dans près de 98 pays et rassemblant près de 107 facteurs de risque, dont le fruit est aujourd'hui exploitable par des équipes de recherche.

D'autres équipes plus aguerries à l'art de la *data science* ont pu mettre au point des algorithmes. Notons chez beaucoup de participants une fraîcheur hardie à s'emparer de sujets requérant parfois la maîtrise d'un socle de connaissances important, le poids de l'autorité sur ces sujets n'agissant plus.

Indiquons qu'à l'occasion de ce Challenge, nous avons pu constater la capacité inclusive du *micro-tasking* car de nombreux participants n'avaient pas les compétences requises tout en se montrant désireux d'apporter leur pierre à l'édifice. Si bien que nombre de propositions nous ont été faites pour proposer à la communauté des tâches plus élémentaires et à portée du tout venant : recherche de jeux de données ouverts, mise en place d'outils méthodologiques, animation de communauté à des fins de recrutement, documentation sur un wiki, *etc*.

// Mega-tasking

Les possibilités du *crowdsourcing* sont vastes, allant du *micro-tasking* à des problèmes complexes que des individus résolvent sans qu'on ne leur en connaisse l'expertise ou la formation. Ces derniers forment là ce que Jimmy Wales appelle « *les experts de leur propre expérience* »[5]. Nous emprunterons à un domaine éloigné de l'épidémiologie un exemple qui illustre un autre aspect du formidable potentiel du *crowdsourcing*, à savoir FoldIt. Il s'agit d'une initiative de l'université de Washington, à Seattle, dont l'idée est l'étude de la dynamique liant la structure des protéines dans l'espace à leurs propriétés fonctionnelles, dynamique encore mal comprise à ce jour. La question étant fort délicate, et étant donnée l'observation de chercheurs selon laquelle la manipulation régulière de ces protéines donne à ses

praticiens une science empirique et intuitive de la façon dont elles se plient sur elles-mêmes, la plicature leur conférant par là-même leur propriété fonctionnelle, des chercheurs ont eu l'idée d'ouvrir un concours et de proposer au tout venant une plateforme en ligne sur laquelle il est demandé de résoudre un problème de plicature des protéines inaccessible au calcul machine. C'est ainsi que des internautes ont montré des facultés à deviner les logiques de conformation tri-dimensionnelle des protéines alors même qu'ils étaient naïfs de toute connaissance en biologie moléculaire. La compréhension de la structure tri-dimensionnelle de la protéase rétrovirale du virus M-PMV, qui fournit un modèle proche du VIH pour tester d'éventuelles molécules inhibitrices, a résisté aux assauts des chercheurs pendant près de dix ans mais a cédé aux efforts des internautes en trois semaines grâce à ce concours. Ceci a fait l'objet d'une publication dans *Nature Structural & Molecular Biology*[6].

Le fait communautaire est un des faits marquant de l'histoire d'Internet, il en est peut-être la nature même. Dans le domaine de la santé, les forums abondent et les communautés de santé sont à l'initiative : qu'il s'agisse de tagger leurs maladies et les effets secondaires des traitements afférents sur des sites tels CureTogether, de mutualiser des données de pollution depuis des objets connectés, de monitorer et partager des paramètres physiologiques dans des communautés dites de *self-quantify*, de cartographier des renseignements de première urgence, tels des défibrillateurs ou des accès pour personnes handicapées sur le site OpenStreetMap, ces dynamiques témoignent de l'émergence d'une intelligence collective, de l'auto-saisine des communautés des problématiques de santé qui les concernent. Ces formes de *crowdsourcing* massives les érigent comme des acteurs à part entière de l'entreprise de construction du savoir médical.

Tout cela illustre la force de l'ouverture de la science à des dynamiques communautaires. Il appartient aux épidémiologistes de mesurer les opportunités d'éventuelles découvertes et de révéler le plein potentiel qui peut naître d'un assemblage hétérodoxe entre l'expert et le profane.

> *L'abondance de données va permettre de nouvelles études épidémiologiques pour définir de nouvelles normes [de nouveaux symptômes qui permettent d'améliorer le diagnostic, ndlr]. »*
>
> ***Dr Jean-François Thébaut***
> *(Usine Digitale, 28/04/16)*

Une épidémiologie globale

// Dans ses méthodes

La recherche bio-médicale est un monde en voie de co-développement. Elle doit intégrer la puissance des logiques distribuées, s'ouvrir par et dans ses méthodes à celles et ceux qui veulent la co-construire. Elle a beaucoup à apprendre, à réutiliser et à se voir proposer. Il s'agit de sortir des logiques de domaine réservé.

L'épidémiologie, augmentée des possibilités qu'offre le *crowdsourcing*, sera amenée à bâtir de nouvelles interactions avec de nouveaux impétrants dans son domaine historique. Elle doit se munir d'interfaces permettant d'inclure des contributions depuis un milieu qui, encore aujourd'hui, correspond à son environnement extérieur.

// Dans ses dimensions

L'épidémiologiste nouveau doit appréhender de nouvelles dimensions : l'animation de communauté, le partage et le co-développement de ses outils avec les citoyens-experts, dans le lien entretenu d'une réciprocité construite. Il doit estimer l'importance de la diversité des enjeux et inscrire sa démarche dans un contexte global, incluant la dimension médicale comme une dimension critique certes, mais en étroite articulation avec un contexte communautaire, juridique et éthique.

// Dans sa co-construction

En effet, dans le sport collectif et décentralisé que deviendrait la recherche en épidémiologie, l'épidémiologiste est aussi *community manager*[7], il sait bâtir un lien de confiance avec des individus dont il est conscient du souci de la protection des données personnelles et il maîtrise les questions de propriété intellectuelle. Il est le pivot, l'animateur d'une communauté et l'intégrateur de contributions dans leur diversité.

Nous noterons que l'épidémiologiste n'a guère besoin de s'égarer loin dans le web pour faire communauté car l'épidémiologie peut tout d'abord s'ouvrir à elle-même et intégrer dans son corpus de méthodes la nécessité de partage de bonnes pratiques,

de documentation de ses méthodes, de veille collaborative et de *problem co-solving*. Des plateformes de *crowdsourcing* telles que Meta Stack Exchange[8] permettent ainsi d'envisager des solutions de co-construction du savoir dans de très nombreux domaines : informatique (Stack Overflow[9]), mathématiques (Mathematics Stack Exchange[10]) et statistiques.

La méta-épidémiologie

Le champ de recherche émergent qu'est la recherche sur la recherche, et qui est sans doute l'un des domaines de la science médicale les plus déterminants, est tout entier tendu vers l'amélioration de la reproductibilité de la recherche. L'effort de nombreuses communautés pour rendre manifestes et porter à l'attention du public les erreurs, les conflits d'intérêt et les fraudes de la recherche bio-médicale, montre une capacité certaine de la société civile à partager et *crowdsourcer* la veille transparente de la littérature scientifique, et à être l'alliée utile, naturelle et spontanée de la recherche clinique.

Conclusion

« *Rien ne se perd, tout se transforme.* »

Loin des mauvais augures qui craindraient la disparition de l'expert au profit du tout-venant citoyen, puis celle du tout-venant au profit de la machine, aucun de ces acteurs ne disparaît ; au contraire, tous sont émergents. En revanche, leurs rôles respectifs évoluent, les compétences se mettent en réseaux, les idées se diffusent bien au-delà des limites dans lesquelles les disciplines les enserrent. Les experts ont un rôle à jouer dans la fabrication du savoir mais aussi dans sa transmission et son interopérabilité dans des formats intelligibles. Ainsi devrait-il en aller de leur nouvelle responsabilité de s'assurer de maintenir une accessibilité et une connectivité maximales aux fruits de leurs savoirs, soit autant d'interfaces entre leur discipline et le monde qui ne lui serait extérieur que depuis l'intérieur.

Il s'agit de consacrer des outils et des méthodes pour assurer la transmission dans un format accessible, qu'il s'agisse de données, de contenus pédagogiques ou d'animation d'une com-

> *Pour faciliter encore ce type de recherche, il faudrait que plus de données soient ouvertes et accessibles et ce, de manière sécurisée et éthique. »*
>
> ***Olivier de Fresnoye***
> *(Up Le Mag, 09-11/16)*

munauté. Ces interfaces sont des points d'articulation essentiels sur lesquels peuvent s'amarrer des communautés plus ou moins informelles de challengers.

Ainsi l'épidémiologie, comme tous les champs du savoir qui ne sont la propriété exclusive de personne, doit être prête à se laisser penser ou modeler par ceux qui n'en sont pas les pratiquants certifiés.

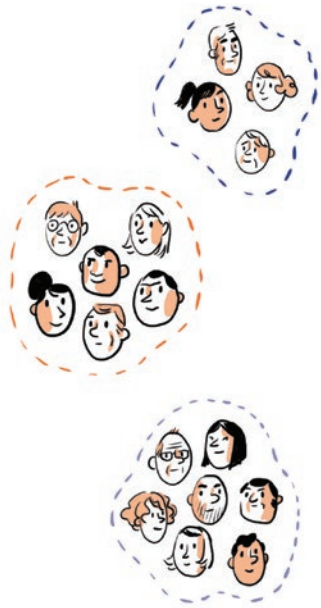
Au total, le maître mot de l'épidémiologie à venir nous paraît être l'ouverture. L'ouverture est un état d'esprit, l'idée qu'une science est par essence ouverte à la réflexion de tous et qu'elle n'est pas une marque déposée. Elle est une pratique, elle est un moyen de s'offrir à un volume beaucoup plus large de propositions et d'efforts. ■

1. Définition Wiktionary <http://fr.wiktionary.org> : « *En linguistique, l'axe paradigmatique concerne le choix des mots eux-mêmes, alors que l'axe syntagmatique concerne le choix de leur placement dans l'énoncé* », dernière consultation le 07 décembre 2016.
2. Faire appel à l'intelligence collective pour traiter d'un objet ou d'une problématique.
3. HealthMap <http://www.healthmap.org/fr/>, développé par des chercheurs, des épidémiologistes et des développeurs au sein du Children's Hospital de Boston, permet de suivre en temps réel le développement et la diffusion d'une maladie en récoltant toutes sortes de données sur le web.
4. Google Flu était une initiative lancée en 2008 par Google qui visait à prévoir les épidémies de grippes à partir des requêtes faites sur son moteur de recherche.
5. Jimmy Wales, "The wisdom of crowds" , The Observer, Londres, 22 juin 2008. Idée selon laquelle nous sommes tous les experts de notre propre expérience.
6. Khatib, F., Di Maio, F., Cooper S., Kazmierczyk M., Gilski M., Krzywda S., ... & Jaskolski M. (2011). "Crystal structure of a monomeric retroviral protease solved by protein folding game players" dans *Nature structural & molecular biology*, 18(10), 1175-1177.
7. Profil qui vise à animer, développer et fédérer une communauté.
8. Meta Stack Exchange <http://meta.stackexchange.com/>.
9. Stack Overflow <http://stackoverflow.com/>.
10. Mathematics Stack Exchange <http://math.stackexchange.com/>.

L'utilité pour le patient et le monde médical

// **AUTEUR**

Muriel Londres



Les bouleversements technologiques impactent potentiellement tous les aspects de l'expérience du patient dans le système de soin, et au-delà même, de l'expérience de vie avec une maladie chronique. Il est important que les malades et les associations les représentant soient associés aux projets et aux recherches permises par cette mutation du système de soins liée à la transformation numérique.

À l'échelle individuelle, les dossiers médicaux digitaux des patients et leur interopérabilité entre les différents lieux de prise en charge doivent permettre d'améliorer celle-ci : plus de sécurité, une meilleure accessibilité, un partage facilité entre le professionnel de santé et le patient, et les professionnels de santé entre eux. Pour la personne vivant avec une maladie chronique, confrontée sans relâche au système de santé - et sans possibilité d'y échapper ! - avoir ses données médicales numérisées

L'utilité pour le patient et le monde médical

doit lui permettre de mieux comprendre sa pathologie, de faire des choix éclairés avec son équipe médicale. À l'heure où il est encore difficile pour le patient de tout comprendre de son dossier médical, où un quart des soins reçus ne serait pas justifié, on peut espérer que des innovations co-construites de nouveaux outils soient facilitatrices.

À l'échelle macro, ces mêmes données regroupées dans des bases de données de plus en plus volumineuses permettront à terme de mieux comprendre les maladies et de mieux les prédire, mais surtout de déterminer les meilleures options en termes de prévention et de traitements.

Face à cette montée en volume des jeux de données disponibles, les associations de patients s'affirment comme acteurs à part entière.

La recherche autour de ces enjeux ne doit pas se faire sans leur participation. Fins observateurs des difficultés de vivre avec la maladie, les associations de patients portent la parole des malades et sont forces de proposition pour améliorer le système de santé. Les données recueillies doivent leur être accessibles, elles sont garantes de l'éthique de leur utilisation et la recherche doit devenir davantage collaborative, avec une prise en compte des problématiques qu'elles mettent en avant.

Epidemium et son Challenge4Cancer ont été inclusifs des associations de patients : nous avons participé au Comité d'éthique indépendant et ainsi au jury. Alors que la présence des malades et de leurs représentants dans certains projets se fait parfois de façon un peu forcée - quand elle se fait -, Epidemium a été demandeur de notre participation et à l'écoute des questions que son Challenge a posées, notamment en termes de finalité d'utilisation des recherches big data et dans une volonté de vulgarisation scientifique. De même, l'aspect communautaire du projet est à souligner. Le Challenge4Cancer s'est déroulé sur plus de six mois et, en parallèle des temps de travail des participants et de leur accompagnement, diverses conférences et moments d'échanges très formateurs ont été organisés. Il a été l'opportunité pour les équipes concourantes et les représentants des comités scientifique et d'éthique de nouer des contacts ainsi que de s'ouvrir aux

idées et aux problématiques des uns et des autres. Au-delà des résultats, qui sont plus prometteurs dans les méthodologies de travail adoptées que véritablement révolutionnaires en termes de finalité des recherches, Epidemium a permis un décloisonnement des acteurs du big data ainsi qu'une réflexion sur une recherche plus collaborative et ouverte. ■

// AUTEUR

Dr Cécile Monteil

En tant que représentante de la communauté Eppocrate, dont l'objectif est l'éveil de la communauté médicale à l'apport des nouvelles technologies et du digital en médecine, j'ai tout de suite accepté de faire partie du Comité d'éthique indépendant lorsque l'équipe d'Epidemium me l'a proposé. Le big data, qui illustre la capacité à collecter et analyser de grands volumes de données, était sur les lèvres de tous mais aucune application concrète en santé n'avait encore vu le jour en France. Il était donc particulièrement intéressant de pouvoir participer à ce programme d'exploration d'un nouveau genre.

Ce qui m'a plu dès le début, c'est l'aspect collaboratif du programme et sa défense de l'*open science*. Le système qui veut que chacun travaille dans son coin représente l'un des grands freins de la recherche en médecine. Le manque de communication et d'interdisciplinarité, l'obsession des brevets et de la propriété intellectuelle, la compétition à la publication ne sont que quelques-uns des obstacles à un travail collectif plus ef-

L'utilité pour le patient et le monde médical

ficace. Epidemium a réussi le défi de rassembler des équipes multi-disciplinaires acceptant de partager leurs découvertes publiquement et en libre d'accès.

En plus de mon investissement pour Eppocrate et de mon poste chez iLumens, le département de simulation en santé de Sorbonne Paris, je travaille à temps partiel aux urgences pédiatriques de l'hôpital Robert Debré. Il faut savoir qu'un médecin n'est jamais amené à rencontrer des ingénieurs, développeurs ou designers, ni lors de son cursus universitaire ni au cours de sa vie professionnelle. Or, on rencontre souvent des médecins ayant de très bonnes idées de création, d'amélioration de dispositifs ou logiciels existants mais qui, faute de savoir comment s'y prendre, ne donnent jamais de suite concrète à ces intuitions de terrain. Combien d'idées intéressantes n'ont ainsi jamais vu le jour ! À l'inverse, combien de gadgets en santé inutiles ou inadaptés sont aujourd'hui sur le marché car pensés par des personnes évoluant hors du monde médical et ne se donnant pas la peine de le consulter.

Pour Epidemium, le rôle du Comité d'éthique indépendant était fondamental. L'utilisation de données de masse provenant d'individus, sains ou malades, est nécessaire pour faire avancer la connaissance mais ne doit pas déroger au respect strict de l'éthique, de la confidentialité et de la protection de la vie privée des personnes. L'objectif était de définir une charte éthique permettant d'encadrer le déroulement du Challenge-4Cancer et des projets créés en son sein, tout en laissant une marge de manœuvre permettant l'émergence d'applications innovantes. Le débat fut riche et mériterait à mon sens d'être mené sur la scène publique. Notre société est aujourd'hui à double personnalité, partageant massivement des données personnelles sur les réseaux sociaux mais barricadant l'utilisation des données de santé.

Les huit projets finalistes nous donnent un aperçu très optimiste de ce que le big data peut apporter à la recherche, notamment dans le domaine de la cancérologie. Nous avons pu voir, grâce à des projets comme CancerViz ou Viz4Cancer, à quel point le mode de visualisation de bases de données est important pour le monde médical dans un contexte de données mas-

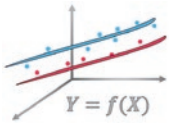
sives afin de pouvoir en extraire des informations pertinentes sans perdre de temps. Ou encore, avec les projets Baseline ou Approches prédictives et risques de cancer, comment des données correctement algorithmées peuvent faire émerger de nouvelles corrélations entre l'apparition de cancer et certains facteurs de risques parfois mal ou méconnus. Le big data est un outil qui va nous permettre de pouvoir travailler plus efficacement dans de nombreux domaines.

Cette première édition fut un succès, autant dans la démarche et la réalisation que dans les perspectives d'avenir. Le big data n'est plus une expression à la mode mais une réalité, un outil qui apporte une valeur ajoutée tangible au chercheur et donc *in fine* au patient. Et au-delà, Epidemium a réussi le pari de démocratiser et de rendre accessible des concepts comme l'*open science* et la recherche collaborative, encore peu connus du grand public : deux concepts qui nous feront aller plus vite et plus loin ! ■

Baseline : modéliser l'incidence et la mortalité du cancer

// **AUTEURS**

Édouard Debonneuil / *Augustin Terlinden*
Peter-Mikhaël Richard

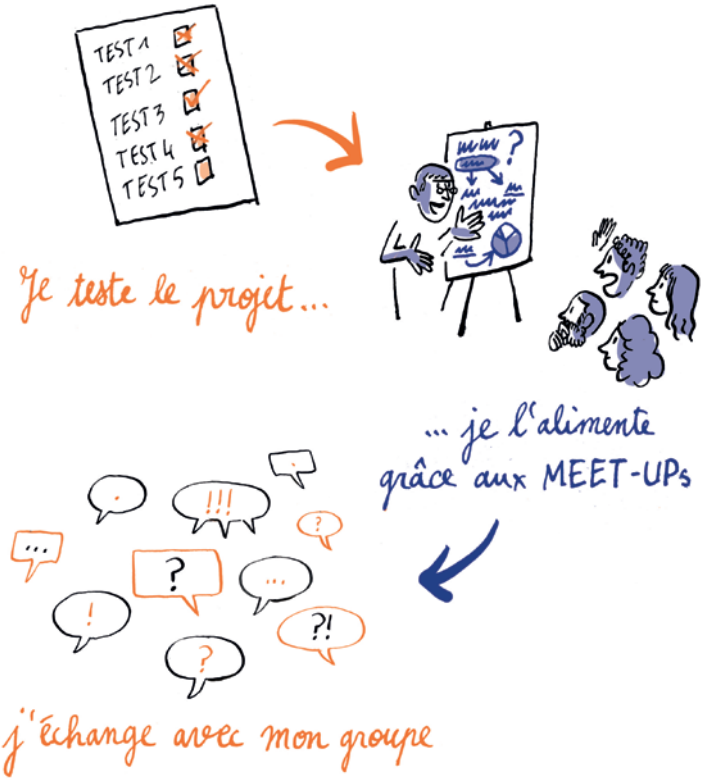


Le projet Baseline vise à prédire l'incidence et la mortalité de divers cancers sur la base de facteurs de risque identifiés dans des données ouvertes de portée mondiale et de granularité régionale. Pourquoi le cancer et pourquoi les données agrégées ? Tout d'abord, les maladies liées au cancer sont mal comprises bien que responsables de millions de morts chaque année. De plus, les données de santé individuelles disponibles sont habituellement rares et les données agrégées sont sous-utilisées, les épidémiologistes leur reprochant, par exemple, les nombreux biais méthodologiques qu'elles peuvent apporter. Ainsi, notre vision était que la solution à un grand problème de société se trouve certainement sous nos yeux !

Dans une première étape, le projet souhaitait élaborer une grande base de données robuste pour la communauté scientifique. Nous avons collecté des données agrégées sur de nombreux sites publics, tels que les *Centers for Disease Control and Prevention* (CDC), l'Organisation Mondiale de la Santé (OMS), l'association Seintinelles <www.seintinelles.com>, *etc*. Ensuite,

en nous basant sur celles-ci, nous avons développé un modèle épidémiologique prédictif visant à extraire des tendances et des liens entre les variables. Ce modèle nous a permis d'investiguer divers facteurs de risque parmi lesquels : la consommation d'alcool, le chômage longue durée, la pression artérielle, le taux de cholestérol, l'âge du mariage pour les femmes ainsi que, pour les hommes, l'appartenance à un groupe ethnique spécifique. Enfin, nous souhaitions procéder à une étape de validation à partir de données individuelles et anonymes mais cela a dû être retardé puis finalement abandonné par contrainte de temps. Nous sommes cependant confiants qu'à terme, ces données serviront à guider la production d'un outil d'aide à la décision pour les médecins généralistes ou encore à conseiller des décideurs politiques sur les allocations budgétaires à réaliser dans certaines aires thérapeutiques.

Le projet a attiré des profils aussi complémentaires que nombreux dont des professionnels de la santé (médecine générale, santé publique, oncologie, épidémiologie), des statisticiens (architectes de données, économistes et actuaires), des développeurs (Web, R / Python, *machine learning* et visualisation de données) et des professionnels de la communication. Tous ont eu leur place et leur chance de participer à ce projet ambitieux. Les enseignements en termes de gestion d'équipe ont été au rendez-vous. Encourager et motiver une équipe de cinquante personnes, en plus de leur activité principale, est un réel challenge. Nous profitons d'ailleurs de ce témoignage pour remercier tous les contributeurs qui nous ont aidés. ■



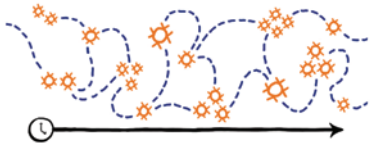
Les fiches

La feuille de route du Challenge4Cancer

2a

2b Les projets du Challenge4Cancer

Thématique 1 : Comprendre la répartition du cancer dans le temps et dans l'espace



Viz4Cancer : site web de visualisation interactive des différents jeux de données disponibles

// Objectif du projet

- Représenter graphiquement l'évolution de différents types de cancers en France et la variation de plusieurs facteur socio-environnementaux.

// Outils

- API Web de traitement de données.
- Visualisation graphique dynamique.

// Qu'est-ce que cela va changer ?

- Permettre aux équipes pluridisciplinaires de disposer d'un langage de visualisation commun, disponible sur le site : **viz4cancer.epidemium.cc**

Wiki : http://wiki.epidemium.cc/wiki/Equipe_Quantmetry

CancerViz : accélérer la phase d'analyse exploratoire de données

// Objectif du projet

- Proposer un outil de data visualisation qui facilite la phase d'acquisition des données et permette d'initier des premières analyses exploratoires.

// Outils

- Une *technology full-stack* basée sur des technologies open-source de traitement, analyse et front-end.

// Qu'est-ce que cela va changer ?

- Un outil de visualisation interactive multicritères accessible sur le site : **cancerviz.weareopensource.me**

Wiki : http://wiki.epidemium.cc/wiki/CancerViz

Thématique 2 : Facteurs de risques et facteurs protecteurs du cancer

Baseline : modéliser l'incidence et la mortalité du cancer selon un grand nombre de facteurs

// Objectif du projet

• La prévention des cancers, en cherchant les secrets qui résident dans les différentes conditions de vie dans le monde.

// Outils

• Techniques (Data Science Studio, MySQL et SQLite...).
• Collaboratifs (hackathons de modélisation).

// Qu'est-ce que cela va changer ?

• Mieux comprendre les facteurs de risques du cancer au regard des conditions de vie pour mieux les éviter demain.

Wiki : http://wiki.epidemium.cc/wiki/Baseline

Approches prédictives et risque de cancer : mesurer l'influence des facteurs environnementaux sur les risques de cancer

// Objectif du projet

• Analyser l'impact des facteurs environnementaux cancérogènes sur l'incidence des cancers.
• Construire un indicateur d'exposition d'une population de référence à certains facteurs.

// Outils

• Codes Python et R.

// Qu'est-ce que cela va changer ?

• Mise en œuvre d'algorithmes de prédiction de l'incidence du cancer.

Wiki : http://wiki.epidemium.cc/wiki/Approches_pr%C3%A9dictives_et_risque_de_cancer

Thématique 3 : Méta-épidemiologie : comprendre le cancer à partir de la littérature scientifique médicale

OncoBase : produire de l'information de qualité sur le cancer pouvant servir de socle commun aux analyses statistiques

// Objectif du projet

- Uniformiser les données hétérogènes disponibles grâce à l'automatisation de la collecte, l'agrégation, l'homogénéisation et l'unification des données.

// Outils

- Programmes d'analyse des données.
- Parallélisation et agrégation.

// Qu'est-ce que cela va changer ?

- Des bases de données issues de la littérature scientifique de qualité plus homogène permettent d'éviter les conclusions erronées et d'améliorer la recherche.

Wiki : http://wiki.epidemium.cc/wiki/OncoBase

BD4Cancer : combiner analyses big data et approches BioNLP pour la pharmacovigilance et la pharmacogénomique

// Objectif du projet

- Identifier les effets indésirables des médicaments anti-cancer.
- Extraire des connaissances à partir de la littérature biomédicale et des essais cliniques pour prédire de nouvelles interactions médicamenteuses.

// Outils

- Environnement d'analyse big data.
- Algorithmes de *machine learning* et librairies NLP.
- Librairies Javascript, ...

// Qu'est-ce que cela va changer ?

- Un système de pharmacovigilance en temps réel et la prédiction de nouvelles interactions médicamenteuses.

Wiki : http://wiki.epidemium.cc/wiki/BD4Cancer

Thématique 4 : Changements environnementaux et cancer

ELSE - Evolutive Life Selection Experience : un jeu éducatif autour d'un personnage né en 2000 qui voit ses risques de cancer fluctuer en fonction des choix qu'il prend

// Objectif du projet

• Sensibiliser aux facteurs de risque liés aux cancers grâce à un outil ludique.

// Outils

• Analyse big data.
• Interface graphique.

// Qu'est-ce que cela va changer ?

• Une application pédagogique mise en ligne et disponible pour tous sur ce site : **conix.fr/epidemium/else.html**

Wiki : http://wiki.epidemium.cc/wiki/ELSE

Venn : avoir une vision globale de la recherche en oncologie environnementale

// Objectif du projet

• À partir des abstracts sur les publications scientifiques de la plateforme Pubmed, extraire et analyser les liens entre cancer et facteurs environnementaux

// Outils

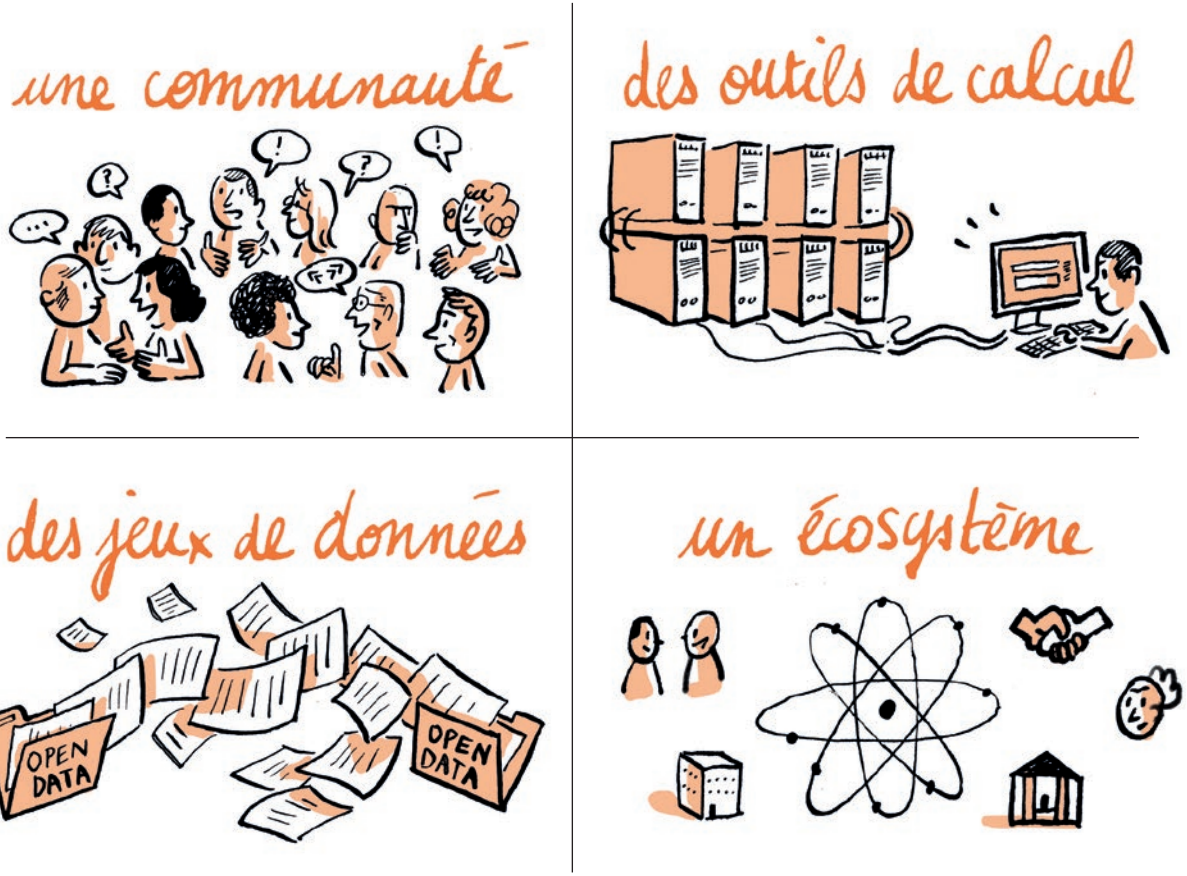
• Outil de recherche intelligent.
• *Machine learning*.
• Logiciel de fouille du texte biomédical.

// Qu'est-ce que cela va changer ?

• Webapp interactive de visualisation des mots-clefs par topic : **venn-epidemium.github.io**

Wiki : http://wiki.epidemium.cc/wiki/Venn

Les ressources du Challenge4Cancer



Pour aller plus loin...

// Quels usages de la science des données et du big data pour la santé ?

- Kawamura T. (2016). “Big data system shows promise in helping cancer patients at Todai”, *The Asahi Shimbun*, 19 septembre 2016, disponible en ligne <http://www.asahi.com/ajw/articles/AJ201609190064.html>, dernière consultation le 30 novembre 2016.
- Schuyler D. (2016). “Is the Blockchain a Potential Cure for Securing Health care data?” dans le site web *Leavitt Partners*, disponible en ligne : <http://leavittpartners.com/2016/09/is-the-blockchain-a-potential-cure-for-securing-health-care-data/>, dernière consultation le 30 novembre 2016.

// Crowdsourcer une épidémiologie du cancer

- HealthMap <http://www.healthmap.org/fr/>, développé par des chercheurs, des épidémiologistes et des développeurs au sein du Children’s hospital de Boston, permet de suivre en temps réel le développement et la diffusion d’une maladie en récoltant toutes sortes de données sur le Web.
- Wales J. (2008). “The wisdom of crowds” dans The Observer, Londres, 22 juin 2008, disponible en ligne <https://www.theguardian.com/commentisfree/2008/jun/22/wikipedia.internet>, dernière consultation le 30 novembre 2016.
- Khatib, F., Di Maio, F., Cooper S., Kazmierczyk M., Gilski M., Krzywda S., ... & Jaskolski M. (2011). “Crystal structure of a monomeric retroviral protease solved by protein folding game players” dans *Nature structural & molecular biology*, 18(10), 1175-1177.

#3

UN CADRE JURIDIQUE ET ÉTHIQUE OUVERT

La réalisation d'Epidemium, programme unique au monde par ses modalités ouvertes et collaboratives, a imposé à ses acteurs de se confronter à des interrogations inédites, tant sur le plan juridique que sur le plan éthique. Parce qu'open science ne signifie pas absence de règles, il s'agit, au-delà du respect de la loi, de structurer un environnement qui garantisse aux participants les conditions du partage de leur travail. Anticiper sur les enjeux éthiques posés par un programme s'attaquant à des données de santé tel qu'Epidemium, a également représenté un formidable défi.

// AUTEURS

Jonathan Keller / Jérôme Béranger / Me David Simhon
Jean-Frédéric Petit-Nivard


éthique
ouverture
collaboration
juridique
open - science
lois

Un règlement pour stimuler le partage et l'ouverture de la science et des données

ARCHITECTURE CONTRACTUELLE RÈGLEMENT

OPEN DONNÉES PERSONNELLES

La présente contribution invite le lecteur à découvrir les raisons de la fixation de règles contractuelles encadrant un programme d'open science. Ce type de recherche novateur repose sur une absence de cadre juridique mais doit comprendre l'accès à des ressources. Il s'agira donc tant de définir les choix contractuels permettant cet accès que de prendre en compte le sort des résultats développés par les participants.

// **AUTEUR**

Jonathan Keller

L'un des tous premiers défis à relever dans le cadre du programme Epidemium était de formaliser un règlement. Il nous a fallu trouver un équilibre entre notre volonté d'offrir un cadre qui stimule le partage et la préservation des droits des créateurs et des propriétaires. En effet, il était essentiel que ce règlement stimule la circulation des connaissances et des données dans un cadre scientifique pour inciter les participants à rejoindre une dynamique communautaire et innovatrice.

La présente contribution s'intéressera plus particulièrement aux différentes solutions proposées pour dépasser les problèmes liés à l'organisation d'une communauté d'*open science*. En effet, en l'absence d'un cadre juridique de référence et en cohérence avec les principes exprimés dans la Charte Epidemium, l'élément contractuel sera privilégié.

Dans le cas spécifique d'Epidemium, nous nous sommes focalisés sur la définition de nouveaux concepts incorporés dans les contrats, comme les frontières de la communauté et de ses membres, ainsi que sur les règles à suivre pour la gestion et le partage des ressources, des informations et des résultats du programme. La suppléance contractuelle au silence légal sera soulignée avant d'aborder la question de la gestion des droits intellectuels.

Quel contrat pour la science collaborative ?

Epidemium se veut un programme réellement ouvert à tous, dans lequel les participants sont libres de choisir leur niveau d'engagement selon leurs disponibilités de temps et leurs compétences, dans un esprit de collaboration et de partage. Le contrat, ou plutôt les contrats imaginés pour satisfaire ce besoin, doivent définir avant tout les limites de la communauté pour octroyer à ses membres les différents droits d'accès et d'exploitation des ressources mises à leur disposition par Epidemium et ses partenaires.

Dans notre tâche, nous n'avons pas pu nous inspirer de la littérature scientifique existante sur l'économie collaborative et les réseaux de création collaborative. Peu d'auteurs ont

cherché à définir ce qu'est une communauté. La jurisprudence ne nous a pas non plus beaucoup aidés puisque la définition d'une communauté n'a pas fait l'objet de décisions judiciaires précises. Enfin, les sciences sociales, qui traitent amplement ce sujet, ont une approche trop éloignée de notre matière pour vraiment être utiles. Nous avons donc procédé d'une façon pragmatique et fonctionnelle en définissant la communauté et ses membres en fonction de leurs besoins et, à partir de cela, les personnes accréditées ont été précisées pour qu'elles accèdent à certains instruments ou outils informatiques fournis par les partenaires du programme.

Définir les membres de la communauté

Dans les projets open source, les plus traditionalistes de la doctrine juridique définissent la communauté comme « tout utilisateur de la création immatérielle ». Vision trop étendue à notre sens car cet utilisateur peut n'être que passif. Sa passivité sous-entend qu'il ne contribue à la ressource commune ni directement en s'insérant dans le projet, ni indirectement en soutenant financièrement ou en promouvant le projet mais en jouissant du résultat final. Nous prônons une vision radicalement différente de la communauté : à notre sens, est membre de la communauté celui qui contribue au projet activement, peu importe que cette contribution soit faite de façon directe ou indirecte et que sa fonction soit une intervention directe ou en support. Les utilisateurs finaux passifs, ou travaillant sur des projets extérieurs au nôtre, ne démontrent pas une volonté de participer au projet, c'est-à-dire une *affectio communitatis*. L'*affectio communitatis* se traduit par la volonté d'adhérer à une communauté.

Afin de faciliter une dynamique communautaire ouverte, souhaitée par les organisateurs, nous avons distingué deux typologies différentes de membres. Nous avons commencé par le « participant » à la compétition Challenge4Cancer. Celui-ci doit s'inscrire sur la plateforme [1]. Il est également tenu d'agréer au règlement Challenge4Cancer et de créer ou rejoindre une équipe pour collaborer à un projet. Ce statut se différencie du statut de « contributeur », auquel il peut toutefois se cumuler.

Ainsi, le « contributeur » se limite à contribuer directement ou indirectement au programme, principalement à travers l'inscription à la partie Wiki du programme[2] pour laquelle il doit accepter les conditions d'utilisation.

Allant au-delà du suivi collaboratif, l'acceptation des différents contrats par le contributeur remplit les obligations légales propres aux hébergeurs, obligations posées par la loi pour la confiance dans l'économie numérique, contraignant l'hébergeur à moduler les commentaires ou à sanctionner les contributeurs le cas échéant.

Outre ce pur aspect préventif, l'identification des contributeurs et/ou des participants confère à ces derniers la possibilité de jouir de certains droits, tels que de participer de manière effective au concours, de pouvoir évidemment le gagner mais surtout d'accéder aux ressources fournies gracieusement par des partenaires tiers.

Un cadre contractuel pour rassurer les partenaires

Certains partenaires du programme ont légitimement émis la crainte que, sous couvert de contribuer à un projet dans le cadre du Challenge4Cancer, le contributeur pourrait utiliser les ressources mises à sa disposition à des fins personnelles et étrangères à ce dernier. Cette crainte est propre à tout projet d'*open science* car ceux-ci, par définition, n'imposent pas de barrages à l'entrée.

Nous avons donc dû prendre en compte cet aspect dans la contractualisation des partenariats en adoptant trois méthodes distinctes et, parfois, cumulables :

- l'allocation de ces ressources est gérée conjointement par les organisateurs du programme et le fournisseur après que le participant a justifié son besoin ;

Un règlement pour stimuler le partage et l'ouverture de la science et des données

- une surveillance opérée par le fournisseur sur les ressources utilisées par le participant ;
- une obligation contractuelle, posée par les stipulations du règlement du concours, à consentir aux contrats de licence d'utilisateur final des ressources mises à disposition.

Maintenant que nous avons établi les grandes lignes des éléments préalables nécessaires à la réalisation et à l'exécution du programme Epidemium dans son ensemble, des éléments connexes, liés à la préparation et à l'exécution du Challenge4Cancer, doivent être à présent étudiés. Ces éléments, propres aux communautés open source, relèvent de la propriété d'autrui mise sous une licence ouverte.

La gestion des atouts immatériels

Le Challenge4Cancer a eu pour principe de permettre aux participants d'utiliser des techniques propres au big data appliquées à l'épidémiologie du cancer, c'est-à-dire de déterminer des tendances et des corrélations par la combinaison de différents jeux de données. Pour que cette combinaison soit possible, encore faut-il que l'information soit licitement accessible, condition qui nécessite à son tour que la donnée ne puisse être qualifiée de donnée personnelle. Au-delà de l'aspect des données se pose également la question de la propriété intellectuelle des développements collaboratifs.

Comment ouvrir l'accès aux données

Dans le cadre du programme Epidemium, l'équipe opérationnelle a fait le choix de constituer un ensemble de bases de données, sur lesquelles allaient travailler les participants, issues de données ouvertes et non « virales ». La viralité correspond à une licence qui contraint les utilisateurs futurs à appliquer la licence d'origine pour toute utilisation et/ou modification future du commun placé sous la licence initiale. Ce choix a été justifié pour plusieurs raisons.

La première raison était de s'affranchir des obstacles contractuels posés par différents producteurs de données. En effet, les contrats de licence de propriété intellectuelle contiennent des

clauses de destination fixant contractuellement la finalité des données, c'est-à-dire la raison de leur utilisation. Or, prévoir et figer la finalité des données, c'est miner le bon déroulement des recherches ouvertes. Rappelons que l'objectif de celles-ci est de parvenir à des résultats scientifiques non connus au départ à partir de ressources communes mises à disposition des participants. Ainsi, et hormis le cadre posé par la Charte Epidemium définie par le Comité d'éthique indépendant, la liberté des participants est absolue.

La seconde raison était que l'accès à des données propriétaires, *a fortiori* contenant des données à caractère personnel, aurait créé une double limitation. Tout d'abord, une limitation de forme. Un contrat de licence classique restreint le nombre de copies autorisées des données. Cette restriction en entraîne une seconde, c'est-à-dire une obligation d'éligibilité de l'utilisateur légitime à y accéder, celui-ci devant être défini pour accéder à certains jeux de données. Or, par définition, l'*open data* est indifférent à la qualité de l'utilisateur final. Cette indifférence devient toute relative si les données ouvertes contiennent des données à caractère personnel, surtout s'il s'agit de données de santé. Le programme Epidemium a pour vocation d'inviter les participants à utiliser uniquement des données ouvertes.

Le partage des résultats développés par les participants

La souscription au règlement du concours dans le cadre du programme est purement volontaire. Cette expression de la volonté n'entraîne pas pour autant de plein droit une renonciation à la propriété intellectuelle en faveur d'une licence ouverte. En effet, le droit de la propriété intellectuelle ne prévoit que de rares cas de cessions automatiques. Ainsi, les participants ont volontairement accepté de contribuer au bien commun en fournissant leurs idées et leurs recherches au sein d'une licence ouverte.

Avant d'entamer la question des différentes licences ouvertes, nous nous devons de préciser qu'il a longuement été question pour les organisateurs du programme de proposer aux participants de contribuer sous licence libre. Néanmoins, ce choix

> *« Invité à présenter les travaux du projet ConSoRe lors d'un meet-up Epidemium, j'ai pu me rendre compte de la force et de l'intérêt de la communauté, à la fois dans le public présent et par les prises de contact que j'ai reçues en parallèle. Aujourd'hui, de plus en plus de personnes comprennent les enjeux des nouvelles technologies et sont avides de savoir comment utiliser ces technologies pour faire bouger la science. Cela, je l'ai découvert chez Epidemium. »*
>
> ***Dr Alain Livartowski***
> *Oncologue, co-directeur des data à l'Institut Curie*

Un règlement pour stimuler le partage et l'ouverture de la science et des données

aurait eu un effet négatif en entraînant un risque de dissuasion psychologique pour les utilisateurs tiers, c'est-à-dire pour toute personne intéressée et néanmoins étrangère au programme Epidemium. En effet, la divergence principale entre le libre et l'ouvert correspond à l'obligation, dans le premier cas, de reverser leur apport au domaine public dans le cas d'une réutilisation des résultats. Or, bien qu'offrant le maintien de la connaissance dans un domaine public artificiel, l'obligation de réciprocité contenue dans les licences libres est vue, à juste titre, comme une atteinte à tout développement ultérieur exclusif des recherche menées.

Enfin, la jurisprudence de la Cour de justice de l'Union européenne a précisé que chaque contenu d'une œuvre de collaboration ou d'une œuvre collective doit être protégé de façon autonome. Un apport qui s'insère dans un projet ouvert doit être fait sous une licence autonome. Par exemple, une datavisualisation d'un projet sera protégée par une licence différente de la licence du texte la commentant ou du code implémentant la solution. Ainsi, en fonction de l'élément protégé, la licence varie. Dans ces différents cas, la licence d'accueil sera ouverte, c'est-à-dire accessible et réutilisable sans autre condition que l'attribution des contributeurs antérieurs. Ainsi pour le logiciel, nous recommandons les licences agréées par l'Open Source Initiative[3] licence Expat/MIT, la licence Berkeley Software Distribution (BSD) ou la licence CECILL-B pour les logiciels ; pour les textes et les images, nous nous contenterons d'inviter les participants à privilégier toute version de la licence Creative Commons Paternité (CC-BY) ; enfin, pour l'ouverture d'une base de données, le choix porterait naturellement sur la quatrième version de la licence Creative Commons avec une obligation de mention de la paternité du producteur ou sur la licence Open Data.

Conclusions

À en juger par la qualité des échanges tout au long du programme Epidemium ainsi que des projets élaborés au sein du Challenge4Cancer, il semblerait que le règlement adopté par la communauté ait accompli sa mission : rassurer tous les

partenaires et les parties prenantes sans pour autant brider la créativité des participants, dans le plein respect des principes exprimés dans la Charte Epidemium *(voir **fiche n°3a**, page 137)* et, bien sûr, des lois en vigueur.

Notre premier souci était de définir les différents profils des membres de la communauté afin de leur octroyer le droit de participer au Challenge4Cancer, de contribuer directement et/ou indirectement au programme, ainsi que de leur accorder l'accès aux ressources mises à leur disposition tout en rassurant les partenaires sur leur utilisation. Par la suite, nous nous sommes attaqués à l'épineuse question des droits des bases de données et de leur utilisation. Enfin, pour accroître les conséquences positives des résultats obtenus dans le cadre de ce programme d'*open science*, les participants se devaient de documenter leur travail, de préciser les algorithmes développés et les données ouvertes utilisées et également de partager les résultats finaux en licence ouverte.

Tout en étant satisfaits des premiers résultats obtenus, nous considérons n'être qu'au début d'un processus qui s'inscrit véritablement dans la durée. Avec le Challenge4Cancer, nous avons en réalité à peine soulevé des questions qui sont à la base de l'ensemble du mouvement *open science*. ■

1. Site de la plateforme Epidemium, <www.epidemium.cc>, dernière consultation le 30 novembre 2016.
2. Site de la plateforme Epidemium, partie Wiki <http://wiki.epidemium.cc/wiki/Accueil>, dernière consultation le 30 novembre 2016.
3. Site Open Source Initiative, The Open Source Definition (Annotated), <https://opensource.org/osd-annotated>, dernière consultation le 30 novembre 2016.

« Il nous a fallu trouver un équilibre entre notre volonté d'offrir un cadre qui stimule le partage tout en veillant à préserver les droits des créateurs et des propriétaires. »

Quelle éthique pour une approche ouverte et communautaire de l'utilisation des big data en santé ?

ACCOUNTABILITY ÉTHIQUE OPEN DATA

NTIC EMPOWERMENT

L'éthique du numérique se traduit par des questionnements, d'une part, sur le comportement et l'usage des individus face au digital, et d'autre part, sur le comportement de plus en plus autonome des outils technologiques en tant que tels. Dès lors, dans le contexte des big data en santé, l'éthique constitue un mode de régulation des comportements basé sur le respect de valeurs humaines que l'on juge essentielles, ainsi que sur un cadre moral pour l'utilisation des données numériques.

// **AUTEUR**

Jérôme Béranger

#3

UN CADRE JURIDIQUE ET ÉTHIQUE OUVERT

Y a-t-il une éthique propre au numérique ? Cette question revient régulièrement et fait débat, tellement il apparaît non naturel d'associer une science humaine à une science technologique que presque tout oppose. Et pourtant, le numérique crée de toute part des injonctions contradictoires qui ont par conséquent des répercussions éthiques spécifiques aux Nouvelles Technologies de l'Information et la Communication (NTIC). Si les big data sont éthiquement neutres, leur usage ne l'est pas, d'où la nécessité pour Epidemium d'établir un cadre et une charte éthique pouvant répondre à tous les projets émergeant dans le programme utilisant les données potentiellement très différentes les unes des autres. En effet, des comportements singuliers naissent des usages de ce nouvel espace-temps que génère le numérique. Les NTIC sont un phénomène culturel voire anthropologique. Ils produisent de nouveaux comportements, de nouvelles visions du monde, et de nouvelles normes sociales.

On peut prendre l'exemple de l'anonymisation qui pose la question de la responsabilité des personnes dont l'invisibilité peut dédouaner de certaines règles de bienséance. L'instantanéité et l'ubiquité que permet l'Internet répercutent nos actes de parole et de pensée de manière conséquente et irréversible. Désormais, l'éthique et la technologie ne doivent plus être mises en relation selon un dispositif à deux étapes. Les questions éthiques doivent faire partie intégrante de leur mission et ainsi construire une réflexion éthique orientée. Dès lors, on ne parle plus d'une approche interdisciplinaire mais plutôt d'une fusion aboutissant à une véritable éthique du numérique où la question des implications sociales et morales s'intègre dans les NTIC.

Dans ces conditions, il devient essentiel d'établir des attentes et des préconisations éthiques spécifiques au monde numérique et de réifier des nouveaux systèmes de valeurs d'éthique et de morale, en gardant toujours en tête cette question : est-ce que le numérique peut induire un risque de mésusage de nos comportements éthiques ?

Depuis quelques années, l'*open data* s'affirme comme un terrain de développement important du big data car les données qu'il

Quelle éthique pour une approche ouverte et communautaire de l'utilisation des big data en santé ?

porte sont considérées comme fiables, intègres, nettoyées de leurs imperfections, ce qui répond à l'enjeu récurrent de qualité des données (Hamel et Marguerit, 2013). C'est la raison d'être des programmes comme Epidemium, qui se sont donnés pour ambition d'utiliser le big data dans un cadre ouvert et partagé afin de produire des résultats de qualité, basés sur la force de la pluridisciplinarité.

Cette pratique consiste à rendre accessible à tous, facilement et gratuitement (sans restriction juridique, technique, ou financière), des données numériques, élaborées par une institution publique ou une collectivité. Cette liberté d'accès et d'ouverture à l'utilisation par des tiers s'intègre donc dans une tendance qui considère l'information publique comme un bien commun dont la diffusion est d'intérêt public et général. La théorie du sociologue américain Robert King Merton met en lumière le bénéfice de l'ouverture des données scientifiques. Chaque chercheur doit apporter sa contribution au bien commun et renoncer aux droits de propriété intellectuelle dans le but de faire avancer la connaissance. C'est la convergence de cette idée scientifique avec les idéaux du logiciel libre et de l'*open source* qui façonne l'*open data* tel qu'il se met en place aujourd'hui. La création de valeur passe davantage par le partage de ces données, leur mise à disposition à des tiers, la participation et la collaboration, que par un effet volume. L'*open data* tente d'introduire un renversement de logique : par défaut, les données et les informations publiques doivent être publiées en ligne – avant même d'être réclamées par des tiers. Ceci marque un bouleversement culturel des mentalités.

Dès lors, l'*open data* est le fruit de son époque où les impératifs de transparence, d'*accountability* sont de plus en plus importants. La transparence est ici perçue comme la réponse à une ère de méfiance, voire de défiance, vis-à-vis des institutions et de leurs représentants. Ce mouvement répond à un ensemble d'enjeux tant économiques que politiques. De l'ouverture des données, on attend des bénéfices démocratiques (meilleure transparence de l'action publique, participation citoyenne, réponse à la crise de confiance vis-à-vis des élus et des institutions) mais également la création de valeur économique par

le développement de nouvelles activités à partir des données ouvertes.

Ainsi, dans le secteur de la recherche et de la santé publique, l'*open data* représente la promesse d'étendre l'indication de médicament, de préserver la santé publique en identifiant les événements massifs ou, au contraire, des signaux faibles comme le début d'épidémie ou des attaques biochimiques, de piloter avec une précision quasi chirurgicale les politiques de santé publique afin de les adapter aux besoins nationaux et locaux de la population, ou d'assurer la sécurité sanitaire en suivant l'utilisation des produits de santé. Dans ces conditions, l'*open data* favorise, d'une part, un meilleur suivi sanitaire à l'échelle d'un pays, et d'autre part, une mise en avant des facteurs de risque à différentes échelles et de travailler sur la détection et le traitement d'épidémies. Sur le plan économique, l'ouverture des données de santé permet une meilleure gestion des dépenses qui passe obligatoirement par la prescription de soins appropriés, sûrs et de qualité. Enfin, les défenseurs de l'*open data* insistent sur l'aspect démocratique de cette pratique, et sur l'importance pour les organismes publics de faire preuve de transparence.

Toutefois, l'*open data* est aujourd'hui confronté à un certain nombre de défis et d'interrogations, tant au niveau de la demande que de l'offre. L'offre de données reste encore largement à construire : la majorité des détenteurs ont ouvert en priorité les données les plus faciles à obtenir tant techniquement, juridiquement que politiquement. Les données numériques perçues comme sensibles, ou celles qui présentent un fort impact social ou sociétal, restent encore largement hors du champ de l'*open data*.

Certaines données sont complexes à appréhender si l'on ne connaît pas le contexte de leur usage premier. N'est-il pas risqué de les rendre accessibles à tous ? Ne va-t-on pas les dénaturer en les interprétant ? En effet, l'appropriation par le plus grand nombre se heurte rapidement à la difficile question de la culture de la donnée. Des compétences d'origines variées sont nécessaires comme : savoir repérer les sources de données, être capable de les traiter, de les manipuler, de porter

#3

UN CADRE JURIDIQUE ET ÉTHIQUE OUVERT

« Outre les résultats et leurs aspects méthodologiques, nous regarderons les questions de gouvernance et d'éthique de l'accès aux données et de diffusion des résultats que poseront certaines des solutions et applications concrètes proposées par les candidats. »

Nicolas de Cordes
Membre du Comité scientifique

Quelle éthique pour une approche ouverte et communautaire de l'utilisation des big data en santé ?

un regard critique sur les conditions de leur production et de leur ouverture, maîtriser les concepts statistiques de base, *etc*.

La confidentialité et les abus dans l'usage de ce type de données ne sont pas à négliger dans ce type de débat. C'est pourquoi on peut redouter une possible « ré-identification indirecte des personnes » en recoupant et en croisant plusieurs informations ainsi que l'utilisation de ces informations par des acteurs privés avec les dérives qui peuvent en découler. Les aspects de sécurisation des données sont donc un élément très important dans l'exploitation de ces *open data*.

Par ailleurs, en éthique, le terme de « valeur » est de l'ordre du devoir-être. C'est un étalon de mesure qui permet de jauger les faits. Il indique des idéaux à poursuivre. Ce mot a une connotation générale et dynamique ; il a d'abord une évocation philosophique avant d'avoir une retombée éthique. Un des fondements de l'éthique est cette impérativité à faire appel à la rationalité des acteurs. Cette idée se structure via une entente dans la coordination, l'échange et le partage entre les protagonistes. Chaque personne contribue à la recherche d'une intercompréhension de la situation à analyser. Cela présuppose donc un certain consensus et une solidarité entre les interlocuteurs qui partagent une même finalité. Si l'éthique est déjà par nature complexe à définir, sa mise en perspective avec le numérique relevait d'un autre défi. L'éthique demande une vision, un dessein, une ambition qui se concrétise dans une orientation.

Aucune technologie ne peut être considérée comme purement instrumentale. Ceci est particulièrement pertinent lorsqu'il s'agit de grands SI automatiques, mis au point pour contribuer à la gestion et l'intégration des grandes organisations, comme les structures de santé. Dans un tel contexte, l'environnement est principalement composé d'êtres humains. En faisant évoluer les SI, les facteurs humains président simplement des facteurs techniques. Même si la satisfaction de ces derniers est obligatoire, ils ne sont jamais vraiment suffisants. Dans tout *big data*, le facteur humain et l'interaction homme-ordinateur sont fondamentaux. Toutefois, dans un contexte multi-utilisateurs simultanés, l'interaction homme-homme est la principale

#3

UN CADRE JURIDIQUE ET ÉTHIQUE OUVERT

Quelle éthique pour une approche ouverte et communautaire de l'utilisation des big data en santé ?

question à résoudre. L'évaluation des grands ensembles de données numériques, tels que ceux trouvés dans le domaine de la santé, est fondée sur le concept de *relationship inter-humaine* (Fessler et Gremy, 2001) qui sous-tendent la conception, la mise en place et l'utilisation des big data. Dans ces conditions, ces « méga données » apparaissent principalement comme un système social, avec ses caractéristiques psychologiques, sociologiques et éthiques. C'est à partir de cette approche que notre réflexion éthique doit commencer sa démarche et poser les principes éthiques spécifiques à l'action numérique.

L'éthique des NTIC peut se découper en trois grandes thématiques :

- l'éthique des données : définissant les principes éthiques garantissant le traitement équitable de données et la protection des droits individuels, tout en utilisant des big data à des fins scientifiques ou commerciales ;
- l'éthique des algorithmes : traduisant l'étude des problèmes éthiques et des responsabilités des concepteurs de données scientifiques, concernant les conséquences imprévues et indésirables, ainsi que les occasions manquées sur la conception et le déploiement d'algorithmes complexes autonomes ;
- l'éthique des pratiques : représentant l'identification d'un cadre éthique approprié pour façonner un code déontologique sur la gouvernance et la gestion des données, favorisant à la fois le progrès de la science des données et la protection des droits des personnes concernées.

Dès lors, la révolution technologique concernant le secteur de l'information doit être menée dans l'intérêt des patients et d'une meilleure prise en charge. En d'autres termes, la seule valeur à prendre en compte, en vue de la conserver, est la personne humaine considérée dans sa dignité d'être moral. Cette dignité humaine constitue une valeur absolue que l'on donne à la personne. Ainsi, des principes éthiques, pratiques, techniques et ergonomiques doivent être imposés afin que les patients et leurs familles restent les principaux bénéficiaires de cette évolution technologique. Cela est d'autant plus vrai que toute réflexion éthique est un conflit entre des valeurs humaines. Quelles que soient nos pensées religieuses, nos

cultures, nos influences politiques ou domaines d'activité, ce sont nos émotions qui révèlent nos valeurs profondes. Comme le souligne si bien Pierre Le Coz (2010) lors de la première journée d'éthique « Cancer et fertilité » de l'Institut Paoli-Calmettes, « s'il n'y a pas d'émotion, il ne peut y avoir de valeurs formalisées et donc pas d'éthique ».

En effet, chaque grand principe éthique peut être associé à une émotion particulière. On peut faire le couplage suivant :

- le respect pour : le principe d'Autonomie (consentement libre et éclairé) ;
- la compassion pour : le principe de Bienfaisance (bien-fondé d'une action) ;
- la crainte pour : le principe de Non-Malfaisance (ne pas nuire à une personne) ;
- l'indignation pour : le principe de Justice (basé sur l'équité et l'égalité).

De plus, dans le cadre de la technologie, l'éthique a affaire à des actes, des actions qui ont une portée causale sociale incomparable en direction de l'avenir et qui s'accompagnent d'un savoir prévisionnel qui, peu importe son caractère incomplet, déborde lui aussi tout ce qu'on a connu autrefois. Elle peut donc se définir comme étant un dispositif de réflexion sur la signification morale de l'action. Cette définition est destinée à être large et fondamentale et à intégrer plusieurs composants de l'éthique informatique (Waskul et Douglass, 1996). On compte principalement cinq applications éthiques relatives aux NTIC :

- l'éthique de l'*empowerment* : associée au patient acteur (e-patient) qui demande son autonomie et sa dignité (respect de ses droits) ;
- l'éthique de l'accès : avec le droit fondamental et la transparence (*Universal Design*);
- l'éthique de la dissémination : relative à une mutation évolutive de l'informatique de contrôle vers l'informatique de service (centralisation et distribution) ;
- l'éthique de la réappropriation : centrée sur les mutations comme potentiels (littératie numérique) ;

« Si les big data sont éthiquement neutres, leur usage ne l'est pas, d'où la nécessité pour Epidemium d'établir un cadre et une charte éthique pouvant répondre à tous les projets émergeant dans le programme utilisant les données potentiellement très différentes les unes des autres. »

Quelle éthique pour une approche ouverte et communautaire de l'utilisation des big data en santé ?

- l'éthique du collaboratif : entourant le partage d'information (sur le Web avec notamment les forums en ligne ou les réseaux sociaux).

Enfin, comme nous venons de le voir, l'information devient le principal objet de l'action morale. Introduire de l'éthique dans le numérique est un acte non naturel du fait que les NTIC seraient dépourvues de toute valeur sociale et humaine. Cette idée provient d'une réflexion commune qui considère que toute technologie est éthiquement neutre, car seul l'être humain peut apporter du sens à ses actions. Pourtant, nous constatons que les big data diffusent également des valeurs dans la mesure où ils impactent et conditionnent la manière dont leurs utilisateurs se comportent. Par conséquent, aucune donnée numérique n'est jamais neutre. C'est pourquoi il ne faut pas réduire l'éthique du numérique à l'expression de valeurs extrinsèques de bons usages de la technologie, mais également aux valeurs intrinsèques à celle-ci. Enfin, avec l'avènement du numérique et des « données massives », c'est toute une éthique qui est à inventer parce que les NTIC dessinent un nouveau paradigme relationnel et sociologique (Doueihi, 2013).

Nous n'avons pas la prétention d'inventer une nouvelle éthique mais plutôt de repenser et réinventer l'éthique existante afin de la faire évoluer vers ce que nous appelons l'« éthique algorithmique » appliquée exclusivement au numérique. Cette nouvelle approche a pour finalité d'intégrer des valeurs et principes éthiques sur la conception, la mise en œuvre, à l'usage des big data notamment dans le domaine de la médecine. ■

La Charte Epidemium : quand l'éthique a vocation à parfaire le droit

// **AUTEUR**

Me David Simhon

Qu'est-ce que l'éthique ?

Soyons fous. Ouvrons un dictionnaire à la lettre E. Pas n'importe lequel. LE Dictionnaire, celui de l'Académie. À la lettre E donc, plus exactement au mot éthique, il est noté, entre autres choses : « science de la morale ». À « moral », dans le même dictionnaire, nous retrouvons « doctrine relative aux mœurs, éthique ». J'imagine volontier l'esprit sagace du lecteur me rétorquer : pas logique ! Pourquoi utiliser deux termes, s'ils sont identiques dans la notion qu'ils recouvrent ? L'éthique et la morale, ce serait la même chose ? Répondons tout net, pour moi (et accessoirement, pour quelques philosophes) : non ! Il y a un monde de différence entre les deux termes.

> *« L'alliance du big data et des questions médicales est extrêmement puissante mais doit en même temps être rigoureusement encadrée par les principes qui ont prévalu à l'exercice de la médecine depuis longtemps. »*
>
> ***Pr Cédric Villani***
> *Membre du Comité d'éthique indépendant*

La Charte Epidemium : quand l'éthique a vocation à parfaire le droit

La morale, c'est le bien et le mal. Avec l'éthique, nous sommes dans le bon et le mauvais.

Exprimé autrement, l'éthique est toujours relative à une époque et un cadre géographique donnés. Il y a 80 ans, en France, l'avortement n'était pas éthique. La peine de mort, si. Aujourd'hui encore, dans certaines sociétés coupées du monde « moderne », il est éthique de manger les cadavres de ses aïeux ! Acte absolument terrifiant, impensable pour les occidentaux que nous sommes ; parfaitement éthique chez ces peuplades.

Par opposition, la morale, on le comprend bien, tend vers l'absolu, vers l'universalité. C'est le « Tu ne tueras point. ». Ce commandement moral **doit** s'appliquer quelle que soit l'époque, les circonstances, la situation géographique.

« Et le droit, là-dedans ? », se questionne le juriste que je suis. La Loi, ce n'est ni de l'éthique, ni de la morale. Des actions immorales ou amorales peuvent être juridiquement autorisées (ne pas honorer ses parents…). À l'inverse, réglementer l'utilisation des pots d'échappement ou la taille des tomates ne relève évidemment pas de la morale. Et surtout, on ne peut pas forcer quelqu'un à être moral, alors qu'on peut le contraindre à respecter la Loi.

À ce petit jeu des définitions, l'éthique se rapproche plus du droit que la morale. Elle peut toutefois, selon les circonstances, dépasser la Loi, être dépassée par elle ou encore lui être complémentaire.

- Sur l'avortement, l'éthique était probablement un peu en avance sur le droit. Avant même 1975 et la loi Veil, l'interruption de grossesse commençait à être tolérée par la société.
- Pour l'homosexualité, il en aura fallu du temps pour que l'évolution des mentalités, l'éthique sociétale, se répercutent dans « notre » droit hexagonal : 1982, dépénalisation de l'homosexualité ; 2015, mariage pudiquement appelé « pour tous ».
- En France, à la fin des années soixante-dix, était-il éthique de décapiter par guillotine un prisonnier ? Probablement.

C'était en tout cas légal et très précisément prévu à l'article 12 du Code pénal : « Tout condamné [à mort] aura la tête tranchée ».

Le 9 octobre 1981 sonna l'abolition de la peine de mort. Je ne suis pas persuadé que l'éthique de nos concitoyens ait pu être totalement bouleversée entre le 8 et le 10. Ce n'est que quelques années plus tard que l'on considéra, communément, la peine de mort comme une sanction non-éthique. Cette fois-ci, la Loi fut plus rapide que l'éthique.

L'éthique a également parfois vocation à parfaire le droit. Avouons-le, dans une approche pratico-pratique, c'est là où elle devient vraiment excitante. L'éthique peut examiner, dans une démarche casuistique, des situations qui ne sont pas forcément ou pas encore envisagées sous l'angle juridique. Là où la règle de droit ne peut pas toujours entrer dans les détails, l'éthique peut la compléter. Ainsi, au nom de l'éthique, vous pouvez vous interdire des actes - l'utilisation de données, pour se refocaliser sur le sujet du *challenge* - qui seraient autorisées par le droit.

Pourquoi penser le réglementaire et l'éthique en amont ?

Sur certains sujets, une fois que vous avez violé la règle de droit ou la norme éthique, vous faites face aux conséquences mais il est déjà trop tard. Les dégâts sont causés, parfois irréversibles.

Si vous deviez récupérer, sans contrôle, ni filtre, les données de santé personnelles et nominatives de l'ensemble de la population française, nous aurions beau dire que cela n'était pas légal, le mal est fait. À l'heure d'Internet et du *cloud*, les données seront potentiellement accessibles par tout un chacun pendant des années, voire des décennies.

Il est important de déterminer en amont ce qui n'est pas acceptable et de s'efforcer de ne pas franchir cette limite. *In fine*, d'adopter une démarche d'anticipation du dommage.

> *L'éthique n'est pas « un perroquet » de la réglementation. Nous ne pouvions évidemment pas aller à l'encontre de la Loi. Mais nous pouvions nous autoriser à aller au-delà. »*

De l'importance de créer un Comité d'éthique pluridisciplinaire

Il est ainsi apparu indispensable de tracer une frontière, avec suffisamment de liberté à l'intérieur de cette borne, pour que les participants puissent travailler et innover. Il fallait une « main bienveillante », celle qui tient mais ne serre pas.

Qui peut se permettre de poser les limites éthiques ? Le chercheur lui-même ? Il serait alors juge et partie, et manquerait immanquablement d'objectivité. Le législateur ou le pouvoir exécutif ? La réponse est forcément négative, car ce n'est alors plus de l'éthique, mais une règle de droit. Le coordinateur du programme ou les partenaires ? Il y aurait là aussi un manque d'objectivité évident. D'où l'idée des organisateurs de faire appel à des personnes qualifiées, des tiers de confiance neutres et indépendants, rassemblées en un comité. Le Comité d'éthique d'Epidemium était né, du moins sur le papier.

Comment le constituer ? Les conventions internationales qui ambitionnent de s'intéresser aux questions bioéthiques insistent sur la nécessité de pluridisciplinarité. Il faut des visions différentes sur un même problème. En France, les comités d'éthique sur la recherche biomédicale, les CPP (Comités de Protection des Personnes), sont structurés en deux collèges (scientifique et non scientifique). À l'intérieur même de ces deux collèges, il y a eu une volonté de rassembler des profils et des formations différentes *(voir encadré ci-contre)*.

Le comité d'éthique créé dans le cadre du programme Epidemium n'a peut-être pas voulu, ou pu, s'organiser aussi précisément dans sa composition. Dans une démarche inverse à celle du législateur (mais pouvait-on attendre autre chose de La Paillasse !) les organisateurs d'Epidemium ont d'abord cherché des personnalités, puis défini les catégories de membres, ... Mais presque par instinct, ils ont souhaité faire appel à des compétences ***différentes*** et finalement ***synergiques*** : le mathématicien, le juriste, le représentant des patients, le praticien, le spécialiste du big data et de l'innovation, l'entrepreneur, l'éthicien, ...

Ce comité s'est révélé aussi peu structuré dans sa manière de se constituer que dans son fonctionnement : pas de président

I. Le premier collège est composé de :

1. *Quatre personnes ayant une qualification et une expérience approfondie en matière de recherche impliquant la personne humaine, dont au moins deux médecins et une personne qualifiée en raison de sa compétence en matière de biostatistique ou d'épidémiologie ;*
2. *Un médecin généraliste ;*
3. *Un pharmacien hospitalier ;*
4. *Un infirmier.*

II. Le deuxième collège est composé de :

1. *Une personne qualifiée en raison de sa compétence à l'égard des questions d'éthique ;*
2. *Un psychologue ;*
3. *Un travailleur social ;*
4. *Deux personnes qualifiées en raison de leur compétence en matière juridique ;*
5. *Deux représentants d'associations agréées des usagers du système de santé.*

(par décision unanime des membres), des discussions très libres, peu de réunions mais de nombreux échanges par courrier électronique. Une sorte « d'auberge espagnole éthique ». Un tohu-bohu organisé, finalement bien ancré dans l'ADN de La Paillasse.

Comme expliqué plus haut, aborder l'éthique dans le cadre du programme Epidemium avait pour intérêt de poser des limites différentes de celles qui peuvent être prescrites par le droit : l'éthique n'est pas « un perroquet » de la réglementation. Nous ne pouvions évidemment pas aller à l'encontre de la Loi. Mais nous pouvions nous autoriser à aller au-delà.

La Charte Epidemium : quand l'éthique a vocation à parfaire le droit

Là où la loi interdirait ou autoriserait, tel un monolithe, l'éthique permet d'encadrer.

Pour prendre un exemple concret, nous avons ainsi été saisis d'une question sur l'utilisation de données ethniques. La loi Informatique et Libertés autorise, sous certaines conditions, le traitement de ces informations.

> *Loi informatique et libertés, article 8 : Il est interdit de collecter ou de traiter des données à caractère personnel qui font apparaître les origines raciales ou ethniques, sauf si la finalité du traitement l'exige pour certaines catégories de données. Notamment les traitements nécessaires à la recherche dans le domaine de la santé peuvent inclure ces informations.*

La règle de droit est connue, ou à tout le moins accessible par qui s'y intéresse. Pour autant, d'un point de vue éthique, doit-on se permettre d'utiliser ces données ? Pas simple. Il est reconnu que sur certaines pathologies, les populations noires sont plus exposées que les populations caucasiennes. Peut-on aborder le traitement des données sous cet angle ?

Nous avons tenté une approche raisonnée et pragmatique : d'accord pour utiliser l'information disponible, mais pas de manière isolée. Nous avons demandé à ce qu'elle soit corrélée à l'environnement et au niveau de vie des populations, afin d'éviter une vision purement biologique, voire eugénique.

Au fur et à mesure des questionnements des uns et des autres, des sollicitations des organisateurs et des participants, le comité a été amené à créer sa propre "jurisprudence". Un corpus de règles que nous pensions suffisamment important pour être écrit. Une charte éthique. Nous avons considéré ces principes comme fondamentaux en 2015-2016 et dans le cadre d'Epidemium. Mais ce qui est vrai en 2016 en France ne l'est pas forcément ailleurs et sera probablement amené à évoluer dans les prochaines années. C'est bien là le drame, comme la beauté, de l'éthique... ■

L'ouverture des données de Roche

// **AUTEUR**

Jean-Frédéric Petit-Nivard

En tant qu'initiateur du projet Epidemium, une initiative éminemment tournée vers l'*open data*, il nous a semblé essentiel d'être pionnier et d'ouvrir nos données pour le bénéfice de la science. Roche France est fier d'être le premier laboratoire pharmaceutique à l'avoir fait en créant Roche Open Database, une base de données ouverte mise à disposition de la recherche sur la plateforme *open data* du programme Epidemium.

Nous vous proposons de découvrir notre cheminement et espérons que ce témoignage pourra vous être utile.

L'origine du projet part d'une conviction forte partagée : l'*open data* est un accélérateur formidable pour la science et plus spécifiquement pour l'épidémiologie du cancer en rendant possible des avancées concrètes sur le cancer pour mieux soigner les patients demain. Avec l'ouverture de nos données dans le cadre du projet Epidemium, notre ambition est de créer un précédent qui fera école en France pour le bénéfice des patients.

Pour réussir Roche Open Database, il nous a fallu constituer une équipe interne regroupant les compétences clés et les expertises d'ordre médical, juridique, réglementaire et analytique. Notre correspondant informatique et liberté a joué un rôle déterminant dans la conduite du projet en maintenant un lien régulier avec la Commission Nationale de l'Informatique et des Libertés (CNIL) dont l'accompagnement et le conseil ont été décisifs pour la réussite de l'initiative.

Avant même de lancer le projet, nous avons d'abord cherché à obtenir une autorisation de la maison mère pour ouvrir nos données Roche France. Le retour a été rapide et positif. Un soulagement car Roche Open Database ne s'inscrit pas dans les modalités prévues par Roche pour le partage des données de santé [1].

L'autorisation obtenue, nous avons pu nous mettre au travail. Schématiquement, le projet peut se scinder en deux grandes parties qui se succèdent : l'une est juridique et l'autre technique.

Aspects juridiques

Pour créer Roche Open Database, la première étape a consisté à bien définir le cadre juridique dans lequel le projet allait s'inscrire.

Ce cadre fait appel à deux notions fondamentales que sont la définition d'un traitement au sens de la CNIL et le consentement patient. La loi définit des règles très claires concernant l'usage des données de santé afin de préserver les intérêts des patients. « *Toute opération […] de collecte, enregistrement, organisation, conservation, modification, extraction, consultation, communication, rapprochement, interconnexion, verrouillage, effacement et destruction* » est considérée comme un traitement par la CNIL.

Pour obtenir la base de données la plus complète possible, notre idée de départ était de regrouper les données de plusieurs études cliniques et de les anonymiser. Or, ce regroupement ainsi que l'opération d'anonymisation des données ainsi regroupées correspondent à un « traitement » au sens de la Loi Informatique et Libertés. Nous avons donc fait une demande d'autorisation auprès de la Commission Nationale de l'Informatique et des Libertés (CNIL).

Nous avons donc fait une demande d'autorisation. Plusieurs éléments devaient y figurer, notamment la finalité du projet, la modalité d'information des patients, les données ciblées, et une évaluation de l'anonymisation sur les critères définis par le G29 (individualisation, corrélation, inférence) [2].

Pour mieux comprendre ces termes, voici une définition succincte donnée par la CNIL [3] :

- **l'individualisation** : c'est la possibilité d'isoler un individu ;
- **la corrélation** : c'est la possibilité de relier entre eux des ensembles de données distincts concernant un même individu ;
- **l'inférence** : c'est la possibilité de déduire de l'information sur un individu.

La demande d'autorisation doit aussi justifier l'article de loi auquel elle fait référence.

Après étude des différents articles, il a semblé que l'article 8 offrait les possibilités les plus adaptées. Nous avons identifié deux pistes possibles : l'anonymisation à « bref délai » et la demande d'anonymisation justifiée par l'intérêt public. Après avoir sollicité une expertise externe, nous avons soumis notre demande de traitement sur la base d'une anonymisation dite à « bref délai ». Ces techniques d'anonymisation s'appliquant davantage à des transactions financières, la CNIL a refusé notre première demande.

Sur sa recommandation, nous avons soumis une nouvelle demande de traitement justifiée par l'intérêt public qui a reçu un feu vert en février 2016. L'autorisation de la CNIL obtenue, nous avons pu lancer la construction de Roche Open DataBase.

Aspects techniques

La construction de la base s'est faite en quatre grandes étapes :

1. Choix des données
2. Transformation et regroupement des données
3. Anonymisation
4. Validation

1. Choix des données

Nous nous sommes concentrés sur l'ensemble des études non interventionnelles finalisées, et réalisées en France depuis

1999 en oncologie. Nous avons ensuite isolé les données d'inclusion [4] de ces études ayant un intérêt pour une recherche épidémiologique sur le cancer.

Après analyse, nous avons retenu douze études représentant environ 8 000 patients.

2. Transformation et regroupement des données

Pour constituer une base regroupant les données de ces douze études, nous avons été confrontés à deux défis : obtenir une structure de base et une nomenclature identiques, et conserver les spécificités des différentes pathologies.

Pour obtenir une structure de base et une nomenclature identiques entre toutes nos études, nous nous sommes appuyés sur les standards définis par le Clinical Data Interchange Standards Consortium (CDISC [5]) qui est la référence pour le stockage de données dans les études cliniques. Ce standard requis pour une soumission à la Food and Drug Administration (FDA [6]) aux Etats-Unis facilite le recoupement et l'exploitation des données cliniques. Cela contribue ainsi à améliorer l'efficience de la recherche clinique.

L'autre défi auquel nous avons été confrontés a été d'intégrer toutes les données dans une structure de base commune tout en conservant les spécificités de chaque pathologie. Dans Roche Open Database, cinq pathologies étaient représentées, à savoir le cancer colorectal, le lymphome folliculaire, les maladies néoplasiques, le cancer du poumon, et le cancer du sein.

De plus, les données collectées dépendent de la finalité de chaque étude, et dans une logique d'éthique et d'efficience, seules les données patients strictement nécessaires à sa finalité sont collectées. Or, il est apparu lors de la constitution de la base que, hormis quelques données standards comme celles ayant trait à la démographie, la plupart des variables étaient spécifiques à chaque étude. Par conséquent, et malgré le regroupement des bases, certaines de ces variables n'étaient renseignées que pour des effectifs de patients relativement faibles.

3. Anonymisation

Après analyse, nous avons identifié deux façons d'anonymiser nos données : l'une conserve la granularité de la base (1 ligne dans la base correspond à 1 patient) en appliquant des techniques de masquage connues alors que la seconde, l'agrégation, regroupe plusieurs données pour former des agrégats (1 ligne dans la base correspond à plusieurs patients).

La première option permet de préserver une plus grande richesse de la donnée. C'est donc sur cette voie que nous nous sommes naturellement engagés au départ.

Pour anonymiser une base de données, il faut d'abord faire disparaître les identifiants directs et ensuite masquer les identifiants indirects.

Voici une définition rapide pour mieux comprendre ces termes :

- **Les identifiants directs** sont les données qui permettent de ré-identifier un individu directement, par exemple le nom/prénom ou le NIR [7].

- **Les identifiants indirects** sont les données qui ne suffisent pas à ré-identifier un individu mais qui, combinées avec d'autres, rendent possible une ré-identification. Par exemple la combinaison (date de naissance, lieu de naissance, code postal, sexe) permettrait de retrouver un individu dans une base.

Dans notre cas, la première étape était déjà réalisée. En effet, toutes les études cliniques sont pseudonymisées, c'est-à-dire que les identifiants directs sont remplacés par une valeur aléatoire.

La deuxième étape, qui consiste à masquer les identifiants indirects, nous a semblé beaucoup plus complexe à mettre en œuvre et les techniques disponibles pour le faire moins robustes. Après analyse, nous avons conclu qu'aucune technique disponible aujourd'hui ne permettrait d'anonymiser parfaitement la base de données [8].

Nous nous sommes donc tournés vers la deuxième option d'anonymisation : l'agrégation.

Cette méthode consiste à regrouper les données de plusieurs patients ayant des caractéristiques communes pour calculer des statistiques (moyenne d'âge, poids moyen, *etc*.). Cette méthode offre un avantage indéniable sur la robustesse de l'anonymisation, même si elle limite les recoupements possibles entre bases.

Pour des raisons évidentes d'éthique, et en accord avec la CNIL, nous avons donc décidé d'anonymiser les données en ayant recours à l'agrégation.

Cette étape nous a permis de générer Roche Open Database, une nouvelle base de données agrégées et anonymisées.

4. Validation

L'étape finale de l'anonymisation consiste à valider la base de données et les résultats obtenus pour déceler d'éventuelles erreurs ou répartitions anormales des données. Nos experts *data analysts* se sont concentrés essentiellement sur deux éléments. Le premier a été de s'assurer que chaque statistique avait été générée à partir d'un nombre suffisamment important de patients, dans notre cas *a minima* dix.

Le second a été de vérifier pour les variables continues qu'il y avait une dispersion suffisante des valeurs pour éviter des cas particulier, notamment de répartition en dirac [9].

Cette vérification marque la finalisation de Roche Open Database. Une base qui allait ensuite être partagée avec tous les participants du Challenge4Cancer d'Epidemium.

Roche Open Database est une initiative véritablement audacieuse qui va dans le sens de la science. En ouvrant ses données sur le cancer, Roche s'inscrit dans la continuité du Plan Cancer qui appelle à rendre plus accessibles les données de santé pour favoriser l'appropriation et l'exploitation de celles-ci par le plus grand nombre.

Le projet n'aurait pas pu aboutir sans le soutien de la CNIL et l'implication d'une équipe experte et pluri-disciplinaire. Au-delà de la création de la base, le résultat principal réside sans doute dans la démonstration de la faisabilité de ce type d'initiatives et nous espérons que ce partage permettra d'encourager d'autres actions similaires. ■

1. Roche partage déjà des données cliniques agrégées avec le grand public via Clinical Trials, un service de l'Institut National de la Santé Americain <www.clinicaltrials.gov>, et des données patients avec les entités de Recherche via Clinical Study Data Request <www.clinicalStudyDataRequest.com>.
2. Le G29, c'est le Groupe de Travail Article 29 sur la Protection des Données des pays membres de l'Union européenne, <http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=50083>, dernière consultation le 30 novembre 2016.
3. « Le G29 publie un avis sur les techniques d'anonymisation », CNIL, article publié le 16 avril 2014, disponible enligne <www.cnil.fr/fr/le-g29-publie-un-avis-sur-les-techniques-danonymisation-0>, dernière consultation le 30 novembre 2016.
4. On appelle « données d'inclusion » les données patients collectées au début d'une étude clinique. Elles peuvent être par exemple démographiques, physiologiques, … : âge, sexe, taille, poids, ...
5. Souza, T., Kush, R., & Evans, J. P. (2007). "Global clinical data interchange standards are here!", Drug discovery today, 12(3), 174-181.
6. La Food and Drug Administration c'est l' Agence américaine des produits alimentaires et médicamenteux, voir <https://fr.wikipedia.org/wiki/Food_and_Drug_Administration>, dernière consultation le 30 novembre 2016.
7. Numéro d'Inscription au Répertoire de l'INSEE, couramment appelé le « numéro de sécurité sociale ».
8. G29, Opinion 05/2014 sur les techniques d'anonymisation, adoptée le 10/04/2014, WP 216, disponible en ligne <http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/index_en.htm>, dernière consultation le 30 novembre 2016.
9. Page Wikipedia de Distribution de Dirac sur <https://fr.wikipedia.org/wiki/Distribution_de_Dirac>, dernière consultation le 30 novembre 2016.

Les fiches

La Charte Epidemium

Les principes éthiques d'Epidemium

2015-2016 - Paris, France

L'accès aux données de masse est une source de progrès importants dans la connaissance des maladies et de leurs déterminants épidémiologiques.

Comme toute innovation, cette utilisation des données anonymes de masse doit se faire dans le respect strict de l'éthique, de la confidentialité, de la protection de la vie privée des personnes, des dispositions légales ou réglementaires en vigueur, et ceci lors des différentes phases du travail : collecte des données qu'elles soient publiques ou pas, analyse, utilisation à des fins d'épidémiologie, d'amélioration des soins, *etc*.

Le comité d'éthique d'Epidemium a pour but de veiller au bon respect de l'éthique tant dans le cadre de la conduite du challenge que des projets qui en seront le fruit. Il énonce ci-devant les grands principes éthiques qui encadreront le déroulement d'Epidemium.

// **Les porteurs de projets devront respecter :**

- Les dispositions légales et réglementaires en vigueur.
- La déclaration des liens d'intérêts.
- La confidentialité et la personne humaine (que les données proviennent d'une source publique ou privée).
- Les règles éthiques préexistantes des données utilisées.
- La sincérité et la transparence dans le recueil, l'analyse et le traitement des données.
- Les principes de bienfaisance et de non-malfaisance au travers d'une évaluation du rapport bénéfice / risque.
- Un engagement à partager à la fois une documentation des travaux ainsi que l'ensemble des résultats et des conclusions qui y sont attachés.

Le comité sera chargé de s'assurer que les projets soumis respectent ces principes.

// **Signataires :** Gilles Babinet, Jérôme Béranger, Emmanuel Didier, Muriel Londres, Dr Cécile Monteil, Pr Bernard Nordlinger, Me David Simhon, Dr Jean-François Thébaut, Pr Cédric Villani.

Le comité éthique se réserve la possibilité de faire évoluer cette charte en fonction de l'évolution d'Epidemium.

Pour aller plus loin...

// Un règlement pour stimuler le partage et l'ouverture de la science et des données

- Bensoussan A. (2016). *Livre blanc : une Science ouverte dans une République numérique*, mars 2016, disponible en ligne <http://www.cnrs.fr/dist/Livre_blanc_DIST_CNRS.html>, dernière consultation le 30 novembre 2016.
- European Commission Directorate-General for Research and Innovation (2016). *Open innovation, open science, open to the world*, disponible en ligne <http://bookshop.europa.eu/en/open-innovation-open-science-open-to-the-world-pbKI0416263/>, dernière consultation le 30 novembre 2016.
- Jean B. (2011). *Option Libre: du bon usage des licences libres*, FramaBook, pp. 307.
- Pelligni F. et Canavet S. (2013). *Droit des logiciels*, PUF, pp. 616.

// Quelle éthique pour une approche ouverte et communautaire de l'utilisation des Big Data en santé ?

- Doueihi M. (2013). *Qu'est-ce que le numérique* ? PUF, pp. 150.
- Ericsson White Paper (2011). *More than 50 Billion Connected Devices*. Ericsson, disponible en ligne <http://www.akos-rs.si/files/Telekomunikacije/Digitalna_agenda/Internetni_protokol_Ipv6/More-than-50-billion-connected-devices.pdf>, dernière consultation le 30 novembre 2016.
- Fessler J-M et Grémy F. (2001). "Ethical problems in health information systems" dans *Methods Inf Med*, 40(4), pp. 359-61.
- Hamel M-P et Marguerit D. (2013). *Analyse des big data : quels usages, quels défis* ? France Stratégie, Note d'analyse n° 8, pp. 1-12, disponible en ligne <http://www.strategie.gouv.fr/publications/analyse-big-data-usages-defis>, dernière consultation le 30 novembre 2016.
- Le Coz P. (2010). « Cancer et fertilité : les aspects éthiques », recherche présentée au colloque de l'Institut Paoli-Calmettes, Marseille, 19 novembre 2010.
- Waskul D. et Douglass M. (1996). "Considering the Electronic Participant: some polemical observations on the ethics of on-line research" dans *The Information Society*, vol.12(2), pp. 129-139.

Conclusion

Gilles Babinet

Utiliser les *data sciences*, les *big data*, pour faire avancer la recherche sur le cancer est, en soi, un modèle de rupture par rapport à ce qui pré-existe. Cette discipline est encore très embryonnaire et son développement passe par beaucoup de tâtonnements et d'itérations.

Toutefois, ce n'est finalement pas en cela que le projet Epidemium est le plus remarquable. Il l'est surtout par sa capacité à proposer un modèle d'innovation ouverte, c'est-à-dire un modèle reposant non plus sur des experts académiquement reconnus, mais sur la multitude la plus vaste possible.

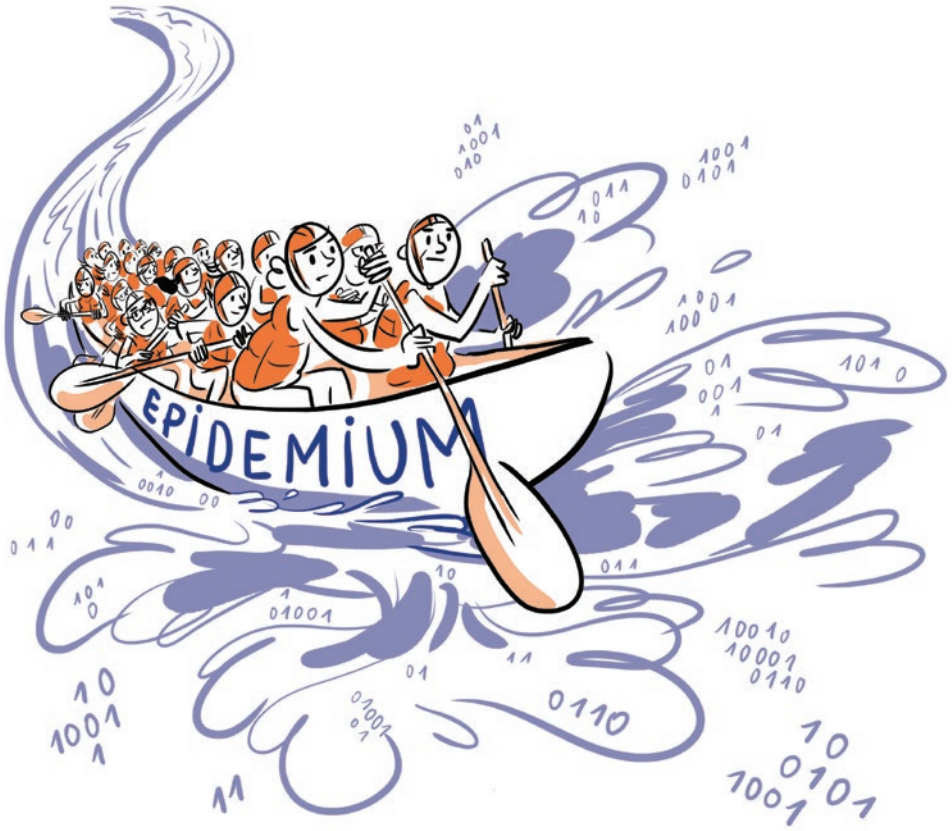
C'est de plus en plus un fait admis : le champ des connaissances scientifiques est désormais si vaste que même les experts ne parviennent plus à embrasser leurs disciplines propres de façon complète. Et de surcroît, dans un monde où la complexité est désormais la norme, et la multidisciplinarité une dynamique fondamentale, il est urgent de changer de modèle d'innovation. Il se pourrait, en effet, que ce ne soit plus au travers des départements de R&D et des centres d'expertises verticales que le futur s'écrive, mais au sein de la multitude.

Qui aurait jamais cru qu'une encyclopédie, Wikipedia, faite par la multitude puisse être cent fois plus volumineuse que la très révérée et référente Britannica ? Et, plus étrange encore, qu'en dehors des biographies, elle puisse contenir moins de fautes que Britannica, ce que des études scientifiques ont démontré ? La multitude, de Wikipedia à Github[1] en passant par Stack Over Flow[2], démontre chaque jour un peu plus sa capacité à être puissante, quantitative, autant que qualitative.

C'est tout le pari de La Paillasse et de Roche à travers Epidemium : créer les conditions pour que tous ceux qui le souhaitent puissent apporter leur contribution pour réussir à envisager des modèles différents. Des modèles qui souvent se retrouvent en rupture avec les modèles communs et académiques.

Néanmoins, pour réussir, il fallait la foi de ses initiateurs, qui n'en ont pas manqué, et avec cette foi, la capacité à gérer de nombreux obstacles réglementaires, liés notamment aux données privées, éthiques et technologiques. Un an et demi après son lancement, on ne peut que constater que ces obstacles ont été brillamment franchis et que le projet Epidemium est désormais sur la piste d'envol.

Plus encore, Epidemium est désormais une source d'inspiration tout à la fois pour les grandes entreprises et les institutions de



tous types. Récemment encore, lors d'un voyage à la Commission européenne, je me surprenais d'entendre un Commissaire y faire directement référence comme le modèle d'innovation qu'il faudrait faire émerger dans le futur.

Il s'agit bien de cela ; si l'Europe, reconnaissons-le, n'est pas aujourd'hui à l'avant-garde de la révolution digitale, elle peut très bien regagner une place enviée, en créant et promouvant le modèle d'innovation de demain. Qu'il s'agisse de la R&D de ses entreprises, d'innovation sociale ou tout bonnement de politiques publiques, il ne fait que peu de doute que l'innovation ouverte dominera un jour sur toute autre forme d'innovation. L'enjeu, pour l'Europe comme pour la France, consiste à saisir ces dynamiques pour parvenir à les accompagner. ■

1. GitHub <https://github.com> est un service web d'hébergement et de gestion de développement de logiciels.
2. Stack Over Flow <http://stackoverflow.com/> est une communauté en ligne de programmeurs pour apprendre ensemble et partager des connaissances.

Liste des auteurs

// **BABINET Gilles, *Membre du Comité d'éthique indépendant d'Epidemium* :** Entrepreneur, *Digital Champion France*.

// **BENBOUZID Djalel, *Membre du Comité scientifique d'Epidemium* :** Docteur en machine learning, post-doc au laboratoire LIP6, Université Pierre et Marie Curie.

// **BENCHOUFI Mehdi, *Équipe coordinatrice Epidemium* :** Médecin de santé publique à l'Hôtel-Dieu, agrégé de mathématiques, Mehdi est fondateur du Club JADE, think tank dédié aux enjeux socio-politiques du numérique (big data, *open culture*, *open science*). Il travaille à des projets collaboratifs de mise au point de technologies médicales en open source.

// **BÉRANGER Jérôme, *Membre du Comité d'éthique indépendant d'Epidemium* :** Chercheur (PhD), Expert scientifique en big data, SI, Ethique et Réglementaire à KEOSYS.

// **BLONDEL Léo :** Doctorant en biologie computationnelle à Harvard, Léo est passionné de sciences. Ayant grandi dans l'univers du hacking et le monde du logiciel libre, il défend la nécessité de libérer la science. « Les cyborgs ont aussi une âme. »

// **CRÉQUIT Perrine (Dr) :** Pneumologue, Méta-analyse en réseau et cancérologie.

// **DEBONNEUIL Edouard, *Membre de l'équipe-projet Baseline* :** Consultant actuariel dans le secteur pharmaceutique.

// **FRESNOYE (de) Olivier, *Équipe coordinatrice Epidemium* :** Spécialiste du développement humanitaire et de la gestion de projet, Olivier a une double formation scientifique et économique. Il a développé de nombreux projets collaboratifs et communautaires innovants et participe à plusieurs groupes de travail sur les nouvelles technologies et l'innovation.

// **FERTÉ Charles (Dr), *Membre du Comité scientifique d'Epidemium* :** Chef de clinique assistant en oncologie à l'Institut Gustave Roussy et expert en bio-informatique.

// **HALDAT (de) Stéphanie :** Directrice de la marque chez Roche.

// **KELLER Jonathan, *Juriste Epidemium* :** Happé par le droit, passionné par les nouvelles technologies et leurs problématiques juridiques, il finalise un doctorat de droit des NTIC à l'Université Paris Ouest La Défense sur la notion d'auteur dans les mondes des logiciels.

// **KOCKLER Leila (Dr), *Membre du Comité scientifique d'Epidemium* :** Représentante Roche, directrice médicale projet à la Direction médicale de Roche France.

// **LANDRAIN Thomas, *Membre du Comité scientifique d'Epidemium* :** Président & co-fondateur de La Paillasse.

// **LÉVY-HEIDMANN Karine, *Équipe coordinatrice Epidemium* :** Community Lead du programme Epidemium, en charge d'animer et de faire grandir la communauté. Elle est également engagée dans l'entrepreneuriat social et membre du conseil d'administration de l'association MakeSense.

// **LONDRES Muriel, *Membre du Comité d'éthique indépendant d'Epidemium* :** E-patiente, coordinatrice adjointe du collectif d'association de malades chroniques [im]Patients, Chroniques & Associés, Militante et bénévole dans l'association « Vivre Sans Thyroïde ».

// **MARIANI Ermete :** Consultant en stratégie de contenu et visualization de la connaissance.

// **MONTEIL Cécile (Dr), *Membre du Comité d'éthique indépendant d'Epidemium* :** Pédiatre urgentiste, Directrice médicale Ad Scientiam et Fondatrice de la communauté Eppocrate.

// **NORDLINGER Bernard (Pr), *Membre du Comité d'éthique indépendant d'Epidemium* :** Service de Chirurgie Digestive et Oncologique à l'Hôpital Ambroise Paré et membre de l'Académie nationale de médecine.

// **PETIT-NIVARD Jean-Frédéric :** Innovation Manager chez Roche.

// **RAVAUD Philippe (Pr), *Membre du Comité scientifique d'Epidemium* :** Professeur d'épidémiologie à l'Université Paris Descartes et à la Columbia University, directeur de recherche INSERM, directeur du Centre de Recherche en Épidémiologie et Statistique Sorbonne Paris Cité, directeur du centre d'épidémiologie clinique de l'Hôtel-Dieu (Paris), directeur de Cochrane Français, directeur du Centre EQUATOR France.

// **RICHARD Peter-Mikhaël, *Membre de l'équipe-projet Baseline* :** Chercheur en thèse dans le domaine de la cosmologie.

// **SIMHON David (Me), *Membre du Comité d'éthique indépendant d'Epidemium* :** Avocat en droit de la santé et Président du Comité de Protection des Personnes Île-de-France III.

// **SANTOLINI Marc :** Chercheur postdoctorant au Center for Complex Network Research de Northeastern University et chercheur affilié à Harvard Medical School à Boston. Ses recherches portent sur la science des réseaux appliquée à la médecine ainsi qu'à l'analyse du travail en équipe dans la production scientifique.

// **TAUVEL-MOCQUET Ozanne, *Équipe coordinatrice Epidemium* :** Diplômée en lettres puis en communication, Ozanne s'intéresse aux nouvelles technologies et aux formes d'organisation qui émergent avec elles ainsi qu'à l'économie collaborative à travers le développement de nouveaux lieux de type Fablabs.

// **TERLINDEN Augustin, *Membre de l'équipe-projet Baseline* :** Innovateur dans le domaine de la santé.

// **VILLANI Cédric (Pr), *Membre du Comité d'éthique indépendant d'Epidemium* :** Mathématicien, Professeur de l'Université de Lyon et directeur de l'Institut Henri Poincaré, il a reçu la médaille Fields en 2010.

Remerciements

Les partenaires

Hypercube, Dataiku, Teralab, Cancer Campus, CapDigital, Global Knowledge, Club JADE, Hacking Health, l'Institut Mines Telecom, Bress Healthcare, SchooLab, Wikimedia France

Les membres de la communauté

Rajaona H., Humbert G., Sourivong F., Ribas A., Martigny P., Clair B., Zhou A., Lafay E., Kamennoff N., Habert B., Frasca A., Bangoula M., Briand F., Leroy G., Betmont C., Iyeze Y., Dunoyer S., Mouhamadsultane A., Duquesnoy G., Agher D., Chatain O., Perez K., Briere T., Boulahya A., Tran T., Maouche S., Hassan M., Verrier J., Bouin O., Bouin G., Baes O., Rouphael R., Agher D., Thibault C., Couralet P., Bohl S., Clément M., Navarro A., Jourdan F., Ben A., Debonneuil E., Terlinden A., Hebert E., Reverdy V., Rouilly V., Richard G., Bouin E., Dejouy V., Strich I., Sportisse C., Yartseva A., Deponge E., Nekooguyan N., Colas E., Tafat Y., Le Clerc S., Noirel J., Coulonges C., Chekroun M., Neveux P., Kerting C., Boyer S., Thea L., Clavairoly A., Mayer C., Sacepe K., Ramdani S., Elisee R., Fagot G., De Rémacle N., Zlaoui K., Foulquié P., Cupillard E., Graveleau M., Lam J., Choffin B., Imbert A., Staron B., Dansokho F., Allorant A., Estival B., Berdugo R., Ravelomanantsoa M., Salgado M., Jandot C., Schreuder N., Foret P., Djian F., Chauvelot L., Le Tiran L., Bereder J., Muaka Di Mavinga G., Hipeaux K., Baiz S., Duge De Bernonville G., Fournier M., Mandelbrojt P., Desrousseaux C., Guinard C., Hamon A., Balmès I., Benbouzid D., Oussar J., Bossi-Malafosse J., Mackosso c., Essayegh H., Burq P., Roy P., Jean-Théodore A., Hadji Z., Letellier P., Picard F., Merinian N., Loménie n., Hadjidj A., Kegl B., Murarasu T., Plaza A., Trinh B., Saysana J., Hachan F., Lopez S., Mukakanamugire Rwagaju A., Ibnouhsein I., Wanono Y., Neuberger K., Jankowski S., Do Cao L., Funtowiez S., Bertrand N., Sutter P., Thiébau N., Jacob Y., Tazi M., Weiss D., Ben Abdallah M., Silbermann T., Gutu M., Zouaoui M., Cinquin L., Mohamed M., Rodrigues R., Dufour P., Bucchini F., Lafay E., Puig Lombardi E., Kejji M., Mezaache S., Rousselle B., Etchebarne A., Kozlowski F., Shinada J., Rafrafi A., Haubensack M., Zouaoui I., Pouchol C., Samelson L., Chera-Piloyan L., Lampe G., Richard P., Bazzoli C., Vincent Q., Zentici J., Gonidec M., Lefevre K., Gangnard J., Gaudin T., Versini J., Fadili M., Leroy G., Choukroun B., Baough L., Mukakanamugire Rwagaju A., Robellaz H., B M., Lara Ramirez A., Zouaoui I., Fabbro A., Le D., Guggiola A., Roux S.,

Leloutre W., Mansar Y., De Hemricourt E., Masquelier M., De Moya j., Kali S., De Chanaud N., Seznec B., Adam E., Ngo F., Adanlété R., Touche C., Denoyer L., Rouyer R., Lombardi J., Sinet L., Pradel M., Motola S., Javed S., Fatni E., Fauconnier J., Roussillon M., Epouhe C., Nadim A., Kpochan N., Taillifet E., Yilmaz J., Landre T., Afane A., Estival B., Blondeau V., Jézéquel G., Boosz P., Louis C., Ngo H., Milhaud X., Tran T., Nguyen T., Nguyen T., Planchet F., Barrau D., Clerget A., Haschka T., Koenig P., Racoceanu D., Do H., Bideault T., Do N., Sarr A., Montagnon A., Naimi G., Aljoufi M., Banquet D., Bouvier J., P J., Zenadi M., Bouguira Z., Pineau H., Vigier N., Shum King M., Callegari V., Barbat L., Messan J., Ferraina M., Feldman S., Petit J., Mutschler A., Demilly A., Meunier G., Panou D., Gea M., Sebastien J., Husson H., Lagasse G., Yang M., Philippe C., Riou T., Bouabdallah C., Giolito A., Scheer J., Scheer J., Bencherif C., Blanchard T., Louhaidia E., Beal M., Brouard C., Rollet R., Fauconnier A., Ziletti A., Le Courtois Du Manoir G., Benoumechiara N., Fischer R., Hilliquin L., El Bachiri A., Dub T., Barbé N., Gimenez U., Bah A., Deneche I., Couderc C., Corteel P., Atameklo K., Warnier M., Gautier M., Giraud P., De Rivet S., Caillé Y., Khalaj g., Caoduro C., Choffe M., Impact A., Sallah K., Salhi S., Schannes B., Litwin S., Bertrand N., Mckeon D., Zidane M., Boudraa F., Ay L., Grob V., Humphery C., Vincent R., Razafindrakoto J., Dadi C., De Vinzelles G., Brisorgueil P., Pinquié R., Berthelot M., Frédéric D., Tanguy A., Vincent B., Donzé L., Mustafa R., Truong T., Dalmasso D., Preau M., Bureau L., Pedro L., Asperti F., Tittmann L., Ould Aklouche K., Zeboulon A., Paix A.

Les partenaires / Têtes de réseaux

Doutriaux R., Thus T., Sakka L., Massart P., d'Ormesson F., Bateson M., Kokshagina O, Sitruk Y., Gruson-Daniels C., Dumas G., Letélier S., Couraud G., Ayache C., Prévost A.-L., Roset S., Sahli A., Taieb D., Walter A.-L., Fayet J., Treguer S., Hélin V., Templier A., Civet A., Bauvin P., Suarez Valencia J. S., Rodriguez A., Kateb D., Klinger E., Jung N., Lods R., Kamennov N., Pitel G., Lefebvre S., Hispa M., Giacopelli M., Teboul D., Mangin O., Delcroix G., Rousseaux A., De Montjoye Y.-A.

L'équipe de Roche Open Data

Magrez D., Violleau M., Lassauge A., Croizat B., Caoduro C., Vlamynck G., Gavini F., Herreye A., Essafi F.

Il était évident que l'analyse de grandes données allait s'attaquer, un jour ou l'autre, à l'un des fléaux qui sévissent encore le plus dramatiquement dans le monde, en fait l'un des plus grands sujets d'inquiétude dans les pays développés : le cancer. Quelle famille, dans notre pays, n'a pas été touchée par cette maladie ? Fléau d'autant plus redoutable qu'il est multiforme, varié, que ses causes et facteurs de risque sont extraordinairement multiples. Et c'est justement pour cela que l'on attend tellement de l'alliance entre grandes données et cancérologie : il y a tant de statistiques mais elles sont si difficiles à interpréter, si variables, que l'on se dit qu'il faudra forcément utiliser des méthodes nouvelles pour en venir à bout et découvrir des choses intéressantes, de nouveaux facteurs que les médecins pourront mettre en œuvre. »

Pr Cédric Villani
Mathématicien - Directeur de l'Institut Henri Poincaré
Médaille Fields 2010

La Paillasse

L'enjeu pour le laboratoire communautaire ouvert La Paillasse est de mutualiser et de distribuer les ressources nécessaires à la concrétisation des projets d'Epidemium : « *À l'ère de l'intelligence collective et décentralisée, il n'y a pas de monopole pour les grandes idées* ».

Epidemium

Epidemium est un programme de recherche scientifique participatif et ouvert dédié à la compréhension du cancer grâce au big data et qui prend la forme d'un data challenge, Challenge4Cancer.

Roche

« *L'étude des mégadonnées ouvertes est un champ de recherche passionnant. En tant qu'acteur de l'innovation en santé, notre ambition est de réinventer une nouvelle forme d'épidémiologie du cancer, pour en faire un véritable outil au service d'une médecine prédictive et préventive* ».

9 791097 214005