Les préparatifs d'un hackathon recherche : au coeur de la fabrique des données

Célya Gruson-Daniel, Constance de Quatrebarbes

Pour citer cet artilce : Gruson-Daniel, C., & de Quatrebarbes, C. (2017). Les préparatifs d'un hackathon recherche : au coeur de la fabrique des données. *Sociologie et sociétés*, *49*(2), 201-221. https://doi.org/10.7202/1054279ar

Les préparatifs d'un *hackathon* recherche : au cœur de la fabrique des données

Gruson-Daniel Célya
Université de Technologie de Compiègne (COSTECH), Université Laval (LabCMO), Centre Virchow-Villermé de Santé Publique Paris-Berlin.
celyagd@gmail.com

De Quatrebarbes Constance
DRISS (*Digital Research in Science & Society)*
4barbes@gmail.com

Résumé :

Depuis les années 2010, de nouveaux formats de recherche tels que les *hackathons* et les *data sprints* se sont développés dans le cadre d'expérimentations en sociologie numérique. Sur un temps très court, ces événements proposent d'analyser des données numériques ou numérisées et d'en présenter les premiers résultats. Or, on observe que ces « formats courts » relèguent souvent dans l'ombre la phase de préparation de ces données pour se concentrer sur l'exploration et la visualisation de jeux de données. En tant que coordinatrices d'un *hackathon* recherche portant sur la consultation République Numérique, nous avons observé les préparatifs à l'œuvre dans l'organisation d'un tel événement. Des observations qui mettent en lumière un important travail de fabrication des données. De leur collecte à leur mise à disposition le jour de l'événement, ces étapes invisibilisées par ces « formats courts » révèlent un ensemble d'enjeux politiques autour de ces *data* et de leur ouverture, qui se dessinent dans les choix mêmes techniques opérés par les acteurs en présence.

Mots-clés : *data sprint*, *hackathons*, *open data*, mises en données, mobilisations politiques

The preparation for an hackathon in research : at the heart of data shaping

Abstract:

Since 2010, new research formats such as *hackathon*s and data sprints have been developed in the field of digital research in sociology. Over a noticeably brief period, these events propose to analyze digital or digitized data and present the initial results. However, we have observed that these "short-form method» often relegate in the shadows the preparation of the data to focus on for the exploration and the visualization of datasets. As the coordinators of a research *hackathon* on the Digital Republic consultation, we observed the preparations required for this kind of event. Observations that highlight the critical work of «data shaping». From the collection of data to its availability for the event, these processes are invisible by these short-form methods. Moreover, this study reveals a set of political stakes around data and open data, which are included in the technical choices made by the actors during this whole process.

.

Mots-clés : data sprint, *hackathon*s, open data, datafication, political mobilisations

Introduction

En sciences humaines et sociales, le terme de recherche numérique est employé pour des projets qui se basent sur le traitement informatisé de données numériques ou numérisées (Plantin & Monnoyer-Smith, 2013). On regroupe, sous ce vocable de méthodes numériques, les techniques et logiciels permettant par exemple la collecte automatique et la fouille de données (*scraping et data mining*), les analyses textuelles et les analyses de réseaux, mais aussi les représentations (carto)graphiques qui y sont associées. Un autre terme employé aujourd'hui, plus générique encore, est celui de *data science* ou science des données, dont l'enjeu est de maîtriser mais également d'interpréter un nombre croissant de données issues des technologies numériques. L'emploi de ces méthodes de recherche s'est accompagné de l'organisation de nouveaux formats de travail à l'image des *hackathons*. Initialement conçus dans l'univers informatique, ce type d'événement s'est peu à peu étendu à d'autres domaines, tout en se diversifiant pour se concentrer plus spécifiquement sur le traitement de jeux de données numériques ou numérisées sur une thématique précise.

Le *hackathon* recherche République Numérique, #HackRepNum, qui constitue le cas d'étude principal de cet article, est une expérimentation de ce nouveau format de travail. Durant une journée, des profils variés[1] - chercheurs, développeurs, juristes,

[1] Dans le cadre de #HackRepNum, des acteurs au profil différent se sont retrouvés lors de cette journée. On comptait des ingénieurs, des développeurs, des chercheurs en SHS (sociologie, sciences de l'information et de la communication, droit), des graphistes, des journalistes, des membres d'associations militantes pour le « libre » et « les communs », ou encore des représentants des partenaires institutionnels et associatifs de cet événement.

membres d'associations, journalistes, designers- se sont regroupés pour analyser les données issues du site de la consultation du projet de loi République Numérique organisé en Septembre-Octobre 2015[2]. Plus de 30 participant.e.s se sont réuni.e.s en six équipes pour analyser les dynamiques de contributions, mais aussi explorer les différents avis émis sur le contenu de chacun des articles du projet de loi. Ce travail s'est conclu par la présentation des résultats de l'analyse de ces *data* à un grand public en fin de journée. En sociologie, le terme de *Data sprints* est souvent employé pour désigner des variantes de ces *hackathons*, centrées spécifiquement sur l'exploration, l'analyse et la visualisation de données. La *Digital Method Initiative (DMI),* regroupant des chercheurs adeptes et/ou curieux de ces méthodes numériques, a ainsi initié à partir de 2012 son premier *Data sprint* lors de sa cinquième école d'Hiver (22-25 Janvier 2013). Le programme de cette formation faisait la part belle au travail en groupe, sous la forme d'une course de trois jours (*sprint*) sur ces *data*. À l'inverse, les sessions plus classiques et individuelles de *keynotes* et *mini-conférences* étaient réduites à une seule journée. L'objectif assumé par les organisateurs du programme, intitulé « *Data Sprint: The New Logistics of Short-form Method » (Digital Methods Initiative, 2013),* était d'expérimenter cette méthode pour format court (*short-form method*)[3]. L'essai semble avoir été apprécié, puisque ce format a depuis été reconduit lors de chaque école d'Hiver. Cette expérience a fait des émules en sociologie. On peut citer par exemple les *data sprints* du programme de recherche en sociologie des controverses EMAPS (*Electronic Maps to Assist Public Science*) (Venturini, Munk, & Meunier, 2016). Le point commun de ces formats, intégrés aujourd'hui dans le champ des recherches numériques, est d'offrir un temps très limité (entre un jour et une semaine) pour analyser les données. Les termes

[2] Ce projet de loi porté par le Ministère de l'Economie et des Finances avait pour ambition de réguler le numérique, ses usages et les nombreux enjeux qui y sont associés (économiques, politiques, sociaux, etc.). La rédaction de cette loi République Numérique s'est insérée dans un temps long avec notamment une phase consultative organisée en tout début du processus législatif (avant même la présentation du texte devant le Conseil D'Etat). Cette phase consultative s'est déroulée du 26 septembre au 18 octobre 2015 sur un site web développé à cette occasion. https://www.republique-numerique.fr/project/projet-de-loi-numerique/consultation/consultation La loi a été promulguée le 7 octobre 2016.

[3] N'ayant pas trouvé de traduction en français de « *short-fort* method », nous avons proposé méthode pour format court. Une autre traduction possible qui nous a été suggérée est celle de méthode accélérée.

propres au régime sportif (sprint et marathon) mettent également l'emphase sur d'autres caractéristiques, comme le *challenge* et le travail d'équipe. Les participants proposent en début d'événement des thématiques sous forme de présentation orale (*pitch*[4]), puis se regroupent en tâchant de rassembler des profils et des compétences variés dans chaque équipe. La métaphore sportive se file autour de l'émulation créée par l'objectif à atteindre : obtenir un prix ou une récompense, ou à défaut, dans tous les cas, présenter son projet aux autres équipes. Pour atteindre ces objectifs, l'organisation est libre et souvent sans programme précis défini, et s'appuie sur des infrastructures numériques pour collaborer en ligne.

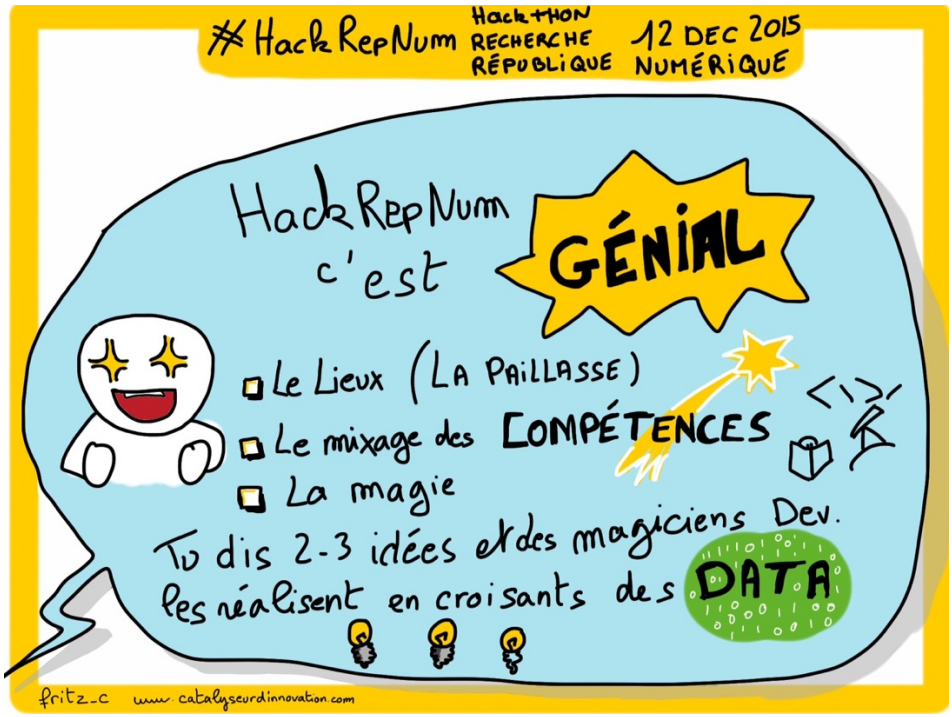

Figure 1 : « HackRepNum, c'est génial ! Tu dis 2-3 idées et des magiciens dev. (Développeurs) les réalisent en croisant des Data ». Cette phrase, extraite d'un dessin réalisé sur le vif par une graphiste lors du hackathon recherche République Numérique, illustre bien l'imaginaire et les enjeux qui entourent aujourd'hui les data. Ici, les développeurs participant à cet événement, par leurs compétences informatiques, peuvent réussir à transformer des données en de « magnifiques » visualisations et en retirer des résultats de recherche.Dessin réalisé par Christelle Fritz lors du hackathon recherche République Numérique (licence CC-BY-SA)

[4] Traduction proposée : *pitch* (argumentaire éclair)

On peut se questionner sur cet essor des formats courts[5] et de leurs usages comme approche de recherche dans le champ des sciences sociales. Que peut apporter à la recherche ces méthodes où les *pitchs* sont mises au cœur de la production de savoirs scientifiques ? Ne serait-ce qu'une mise en scène au détriment de la production de connaissances de qualité ? Mais surtout, que peut-on produire sur des temps si courts ? Les sociologues qui participent à ces événements ne plongent-ils pas tête baissée dans cette utopie technicienne revisitée à l'ère numérique (Proulx, 1984), en surévaluant l'importance des données et de leur maîtrise ? Malgré les doutes et critiques qui peuvent être soulevés, on remarque cependant que ces *data sprints* et *hackathons* perdurent dans ces nouvelles branches de recherche numérique. Cet article vise à interroger l'intégration de ces modalités de travail en sciences sociales, en évitant le double écueil de faire l'éloge de ces nouveaux modes de recherche et d'extrapoler sur l'extinction de toute réflexion sociologique. Nous souhaitons plutôt étudier ce qu'induisent en pratique ces formats courts, en soulignant certains enjeux d'ordre politique qu'ils peuvent soulever[6]. C'est plus particulièrement le temps limité de ces formats que nous souhaitons questionner, en étudiant le travail qu'opère de tels *hackathons ou data sprints* sur les données numériques.

Pour répondre à ces questions, nous nous appuyons sur des travaux théoriques issus des études sur les sciences et les technologies (STS), qui de longue date ont montré l'important travail de construction des données et des faits scientifiques (Gitelman, 2013; Star & Griesemer, 1989). Plus récemment, avec l'usage croissant de ces technologies numériques, des travaux se sont plus spécifiquement concentrés sur les problématiques de stockage, de collecte et d'analyse dû à la multiplication de ces données numériques dans des secteurs de plus en plus nombreux. Plusieurs travaux ont portés sur les processus sociotechniques accompagnant cette « fabrique des

[5] Pour désigner par la suite l'ensemble de ces événements (*hackathon, sprint et camp*) en recherche, nous emploierons le terme plus général de formats courts, pour reprendre la notion de *short-form* employée par la *Digital Methods Initiative* ou bien de *short course* pour le *Humanities hackathon* de 2012, cité en introduction.

[6] Notre réflexion s'inscrit ainsi dans une réflexion de sociologie générale. Malgré les glissements lexicaux utilisés (de données à *data*, de statistiques à *data analysis*, de communication électronique au substantif « numérique »), notre objectif est d'apporter un regard réflexif sur les évolutions que le numérique apporte sur les pratiques même du sociologue.

données » [7]t et leur mise à disposition dans le domaine des sciences mais aussi de l'administration publique et des entreprises (Denis & Goëta, 2017; Mabi & Plantin, 2017; Menger & Paye, 2017). Ces études soulignent souvent l'invisibilisation même de ces étapes au cours desquelles les données sont façonnées et mettent en lumière les tensions et négociations à l'œuvre, qu'implique leur maîtrise et leur diffusion (Dagiral & Parasie, 2017; Denis & Goëta, 2017) . Ces études soulignent aussi quelques spécificités propres au « numérique », en montrant comment les dispositifs numériques sont eux-mêmes représentatifs de nouvelles thématiques politiques associées à la défense de la libre circulation de l'information mais aussi à de nouvelles formes d'actions politiques et de mobilisations qui s'ancrent dans des choix techniques (Briatte & Goëta, 2014; Coleman, 2011; Granjon, 2015; Mabi & Plantin, 2017).

À l'appui de l'étude de cas sur le *hackathon* recherche République Numérique, nous souhaitons mettre à jour une série d'opérations socio-techniques spécifiques de la préparation des données numériques et de leur mise à disposition aux participant.e.s dans le cadre des formats courts. À la différence d'études antérieures qui portaient sur les caractéristiques de ces événements et l'intérêt de ces formats pour analyser ces données numériques de façon collective[8] , nous opérons un pas de côté en nous concentrant plus particulièrement sur les étapes situées en amont de ces événements, souvent invisibilisées au profit de regards concentrés sur le déroulement de l'événement lui-même. En tant que coordinatrices de cette journée, nous avons pris part aux préparatifs de l'événement aux côtés d'un ensemble d'acteurs et de partenaires. Notre analyse repose sur ce cas d'étude qui fait ressortir divers enjeux socio-politiques encapsulés dans les choix techniques qu'opèrent les acteurs en présence. Quoique d'un format court, nous avançons que ces *hackathons* sont à considérer dans un temps plus long. Par la mise à disposition de jeux de données préalablement collectées, ces événements participent à l'invisibilisation de la « fabrique des données ».

[7] Nous employons le terme de « fabrique des données » en référence aux travaux de S. Goëta et J. Denis dont les travaux ont porté sur les coulisses de l'open data et la mise en évidence par une approche ethnographique d'un processus de « brutification » des données qu'ils nomment la « fabrique des données brutes » au sein des administrations publiques en charge de l'*open data* (ouverture des données).

[8] Nous donnons par la suite quelques exemples de ces évènements.

Dans la suite de cet article, nous allons tout d'abord revenir sur quelques éléments théoriques issus des STS sur la fabrique des données. Nous les croiserons avec un ensemble d'études plus récentes associées aux spécificités de la maîtrise des données numériques et des nouvelles problématiques qu'elles suscitent. Nous reviendrons ensuite sur les origines de ces *hackathon*s et leur adaptation dans d'autres domaines, notamment sous la forme de *data sprint* qui tend à reléguer en amont des événements la préparation de ces jeux de données. Ces étapes préliminaires, en amont du *hackathon* recherche République Numérique, constitueront l'objet principal de notre analyse. Après avoir précisé le contexte général de cet événement, de son organisation et les raisons de l'intérêt qu'il a suscité, nous détaillerons le sens pris par la « mise en données » en montrant qu'il s'agit à la fois d'un travail d'extraction mais aussi de mise à disposition et d'ouverture de ces jeux de données sous un format impliquant des « mises en vue » particulières. Seront ainsi dévoilés un ensemble de choix techniques opérées par les acteurs, qui incorporent/contiennent des formes d'actions politiques d'un nouveau genre.

1- La « fabrique des données » et leur invisibilisation

Les études en STS ont apporté un éclairage sur le rôle des infrastructures et de la « matérialité » des artefacts dans le champ scientifique et dans la production des connaissances (Bowker & Star, 2000; Star & Griesemer, 1989). Par des démarches souvent ethnographiques au sein des milieux scientifiques, ces études ont détaillé le travail nécessaire à la constitution des données et y ont analysé l'organisation des équipes de recherche (Flichy, 2013; Heaton & Millerand, 2013; Millerand, 2011). Que ce soit dans le champ de la médecine, de l'écologie, de la botanique, ces travaux ont montré comment la constitution même des bases de données consistait en une série successive d'opérations de façonnage, d'étiquetage, de réduction, de standardisation pour passer d'un objet complexe - un ensemble de faits observables ou collectés par des instruments scientifiques - à la notion même de données manipulables et calculables . (Dagiral & Parasie, 2017; Dagiral & Peerbaye, 2012; Heaton & Millerand, 2013) Ces études ont aussi mis en lumière le rôle de travailleurs « invisibles » sur les modalités de représentation de ces données, et plus globalement de production de

savoirs (Mauz & Granjou, 2011). Le domaine bio-médical et des sciences du vivant a souvent été le terrain d'étude privilégié de ces travaux empiriques. Ces dernières décennies, ces domaines ont connu des modifications de leurs modes de travail et d'organisation, suite au développement de nouvelles infrastructures et d'une masse d'informations de plus en plus importante à analyser (Dagiral & Peerbaye, 2012). Le champ des administrations, avec le développement des grandes enquêtes et l'utilisation des statistiques publiques (Desrosières, 2008), a aussi été l'objet d'étude de cette fabrique des données et des enjeux de gouvernance qui lui sont associés. Ces travaux rappellent qu'il n'a pas fallu attendre le « numérique » pour que l'on porte attention au façonnage des données dans un réseau socio-technique.

On note cependant une multiplication de ces problématiques et une amplification de ces enjeux avec le développement des technologies numériques dans de plus en plus de secteurs. Par *data* aujourd'hui, on entend la plupart du temps des données nativement numériques (sans passer par une étape de numérisation). Ces données sont produites par la mise en place de dispositifs numériques (site web, applications mobiles, etc.) et de l'usage que l'on en fait aussi bien dans le cadre professionnel, personnel et dans bon nombre de nos démarches quotidiennes (santé, administration, etc.). Nos activités politiques peuvent également se voir « transformées en données », à l'image de la consultation République Numérique. La participation à cette consultation en ligne est répertorié par un site ou une plateforme et stocké dans leur base de données. Face à la production croissante de ces données (Boullier, 2016) bon nombre de secteurs déploient des méthodes de traitement informatisé et automatisé de l'information afin de maîtriser ces données et d'en tirer du sens. Ces *data* constituent ainsi un horizon d'attentes fort (Loveluck, 2015; Turner, 2012), mais aussi de nouvelles formes de revendications. Les qualificatifs associés à ces data aujourd'hui en révèlent quelques éléments clefs. D'un point de vue technique mais aussi économique, l'emploi du terme *«Big Data»* met en exergue les enjeux du traitement de données hétérogènes, produites en permanence via Internet, nécessitant des méthodes de stockage, mais aussi de requêtage, filtrage et d'analyse spécifique. Le terme de *data science* traduit en français par science des données a été employé depuis les années 2000 pour désigner le travail opéré sur ces data, nécessitant des compétences à la croisée entre statistiques et informatique ainsi que

la maitrise des logiciels associés (Dagiral & Parasie, 2017). Le qualificatif d'*open* quant à lui souligne l'apparition de nouveaux enjeux démocratiques autour de la mise à disposition à toutes et tous des données publiques. Cet *open data* devient le fer de lance de revendications politiques concernant la transparence des pouvoirs publics et un moyen d'*empowerment* des citoyens.

Des études à la croisée entre STS, sciences de l'information et de la communication (SIC) et sociologie portent aujourd'hui leur attention sur cette fabrique des données numériques en considérant la «matérialité» même des dispositifs numériques (Bigot & Mabi, 2017; Mabi, 2013). Des études se sont ainsi attachées à étudier « ce que les *data* font faire aux SHS (et vice versa) » (Jaton & Vinck, 2016). Elles se sont portées notamment sur le cadre socio-technique qui entoure la production, la diffusion et la valorisation de ces données. Ces travaux montrent d'une part que la manipulation de *data* en SHS mène à une collectivisation des sciences sociales, c'est-à-dire un travail de plus en plus collectif entre des profils différents, et d'autre part que ces rencontres ne se réalisent pas sans tensions, réticences et parfois incompréhensions (Jaton & Vinck, 2016). Le terme de *datafication* ou de « mise en données » est employé aujourd'hui pour désigner cette « fabrication des matériaux à partir desquels des traitements statistiques et des analyses se déploient » (Jaton & Vinck, 2016). Cette fabrication consiste notamment à rendre « brutes » ces données (Gitelman, 2013). L'étude de la constitution de ces « sources brutes » dans le cas de l'ouverture des données publiques (*open data*) (Denis & Goëta, 2014; Denis & Goeta, 2013) a mis en avant le travail effectué par un ensemble d'acteurs des administrations pour « ouvrir » ces données. Ces mêmes administrations sont aussi sujettes à des réorganisations institutionnelles mais aussi à une modification de leurs infrastructures avec le développement de services et de portails *open data* pour soutenir l'idée d'un gouvernement ouvert (*open gov*) (Schrock, 2016). Ces activités sont ainsi un écho intéressant à l'important travail de petites mains en sciences biomédicales (Mauz & Granjou, 2011), qui rend souvent invisible les enjeux techniques et socio-politiques entourant la construction – mais aussi la mise à disposition – des jeux de données[9].

[9] On aurait avec cette institutionnalisation du « hack » une récupération politique et marketing de ces formats devenus formules magiques de l'innovation numérique. L'étude de (Zukin & Papadantonakis,

Ces travaux soulignent également certaines spécificités propres au numérique, notamment dans les formes et objets de revendications politiques. Internet et les dispositifs numériques sont en effet aujourd'hui des objets de mobilisations en tant que tels (Coleman, 2011, 2012). Ces revendications ont débuté tout d'abord dans le domaine informatique autour des logiciels libres (*free software*) dans les années 1990, pour défendre un libre accès mais aussi une libre réutilisation du code source informatique face au développement de logiciels prioritaires (Broca, 2013). Avec le déploiement de ces technologies numériques, ces revendications se sont étendues à d'autres domaines autour d'un dénominateur commun, celui de l'information et de sa libre circulation. Cette information peut consister aussi bien en du code informatique (logiciels libres), un article scientifique (*open access*), mais aussi, comme nous l'avons vu, des données publiques ou d'intérêt général (*open data*). Aujourd'hui, la notion de « communs informationnels » fédère divers acteurs qui défendent le développement de modèles de gouvernance communautaire autour de ces ressources numériques (Broca, 2013; Peugeot, 2014). À ces objets politiques, se sont aussi ajoutées des formes spécifiques de mobilisations et d'action politiques. (Granjon, 2017; Gruson-Daniel & Mabi, 2017). Qualifiées parfois de techno-pragmatiques, celles-ci s'ancrent dans des répertoires d'actions aux fortes composantes techniques et juridiques (Cardon & Granjon, 2013; Granjon, 2015). Elles puisent notamment dans une éthique du *hack* qui met en avant l'expérimentation mais aussi la créativité dans l'optique de contourner des problèmes existants[10] (Broca, 2013; Coleman, 2014; Granjon, 2017; Loveluck, 2015). Les revendications portées par les acteurs qui défendent l'idée d'un *open data* concernent notamment les formats des jeux de données. Des formats ouverts tels que le .csv au lieu d'un .xls représentent ainsi des enjeux pour assurer l'accessibilité mais aussi l'interopérabilité de ces informations et pour en faciliter la réutilisation et la circulation sans dépendre de logiciels propriétaires (Denis & Goëta, 2017; Goeta, 2016).

2017) sur les hackathons soulignent de nouveaux modèles économiques basés sur l'exploitation de ces données par les participants lors de ces évènements.

[10] Le terme d'hacktivisme a été notamment employé pour décrire plus spécifiquement des formes d'expression contestaire qui se fondent sur ces modalités informatiques et du hack (Granjon, 2017).

La mise en évidence des enjeux – notamment politiques – qui entourent la fabrique des données et leur partage nous conduit à questionner le développement des *hackathons* mais aussi des *data sprints* dans le cadre de ces recherches numériques en sciences sociales. Nous allons donc revenir brièvement sur l'origine de ces formats issus du milieu informatique, et sur leur adoption dans de nombreux domaines. Une adoption accompagnée d'une adaptation, à l'image des *data sprints* qui, en amont des événements, a bien souvent invisibilisé le travail de préparation des données.

Les termes de *hackathon* et de *sprint* émergent dans le milieu informatique nord-américain à la fin des années 1990-début des années 2000[11]. L'emploi du terme *hackathon* s'est ensuite généralisé, dès la fin des années 2000, aussi bien du côté des hackers que des entreprises logicielles désireuses d'innover. La notion de *sprint,* avant de se rapporter à un type d'événement comme dans le cas du *data sprint*, renvoyait à un élément central des méthodologies agiles issues du développement informatique. Développées dans les années 2000, elles reposent sur une approche résolument empirique, consistant en une série de cycles de développement de courte durée désignés par le terme *sprints* (Sutherland & Schwaber, 2013). Elles ont été déployés dans le milieu des entreprises comme le mode d'organisation (collaboration distribuée) le plus adapté pour développer des outils informatiques non propriétaires désignés sous le terme de logiciels libre et *open source* (Broca, 2013). Au fur et à mesure du déploiement des technologies numériques dans d'autres domaines, ces événements se sont étendus à d'autres secteurs. À partir de 2010, ils se sont éloignés des objectifs technologiques initiaux (améliorer un logiciel, développer une nouvelle application, etc.), pour valoriser une thématique (la résolution d'une problématique sociale, du marketing) ou pour cibler une population particulière de participant.e.s (les femmes par exemple) (Briscoe & Mulligan, 2014). Certains de ces « formats courts »

[11] Il faut remonter à l'année 1999 pour voir apparaître les premiers *hackathons* au sein de deux communautés différentes, d'une part dans le cadre d'un projet de logiciels libres (Open BSD) et d'autre part lors d'une conférence Java One organisée par l'entreprise Sun microsystem. Les *hackathon*s d'Open BSD, qui se déroulent encore aujourd'hui une fois par an, mettent en avant le « *shut up and hack* ». Le but de ce premier *hackathon* était de développer et d'intégrer à la fin de la semaine une fonctionnalité de cryptage des données à leur système d'exploitation (*operating system*), cela afin d'éviter des problèmes juridiques liés au déploiement de logiciels de cryptographie américains (Briscoe & Mulligan, 2014). On constate déjà ainsi l'importance des enjeux juridiques associés au développement de tels projets. L'autre *hackathon* consistait en une compétition pour écrire un programme en Java au sein même de la Conférence.

se sont notamment concentrés sur l'analyse de jeux de données mis à disposition de participants, comme dans le *Hacking health* dans le domaine de la santé (HackEbola with Data, 2015). Notons aussi, dans le domaine de la culture, un premier *hackathon* organisé par la Bibliothèque Nationale de France en novembre 2016, pour permettre « 24h d'émulation et d'échanges » et « imaginer ensemble la bibliothèque du futur »[12] avec la réutilisation de données publiques. Les *Data sprints* organisés en sociologie se situent dans la même veine.

Dans ces événements, l'étape de préparation des données, que l'on nomme aussi « mise en données » ou *datafication*[13], n'est souvent pas visible . Il s'agit, pour les participants, de se concentrer sur la phase d'exploration, de traitement, d'analyse et de visualisation des jeux de données déjà disponibles[14].En sciences des données, ces étapes de collecte et de préparation des données sont connues pour être un long processus, chronophage et fastidieux, qui nécessite de collecter, nettoyer les *data* et de les standardiser pour produire des jeux de données analysables par la suite par différentes méthodes statistiques et algorithmiques[15]. Or, on peut se questionner sur ces étapes préliminaires effectuées en amont de la tenue de de ces évènements Quelles données sont mises à disposition des participants ? D'où proviennent-elles ? Quels traitements ont-ils été opérés – et par qui l'ont-ils été – pour les rendre disponibles aux participants ? Ces formats courts ne participent-ils pas à invisibiliser

[12] Un exemple plus récent est celui d'un *data sprint* organisé sur les données du fond national d'art contemporain avec pour but de « travailler ensemble sur cette collection particulièrement riche, et tester des hypothèses de recherche, par l'exploration et la visualisation d'informations. » https://www.inha.fr/fr/agenda/parcourir-par-annee/en-2017/novembre-2017/un-datasprint-sur-les-donnees-du-fonds-national-d-art-contemporain-nouvelle-page.html

[13] Nous reprenons ici la définition de Bastin et Francony pour la datafication comme « la fabrication des matériaux à partir desquels sont produits statistiques et jugements de faits, a nettement moins été discuté. » (Bastin & Francony, 2016).

[14] . Ces jeux de données peuvent provenir par exemple de portails *open data* ou bien d'un travail préliminaire d'extraction d'informations d'un site web ou des réseaux sociaux numériques.

[15] Les méthodes d'analyse de ces données reposent aujourd'hui sur des traitements computationnels réalisables par la puissance de calcul des outils informatiques. Ces méthodes associent à la fois des savoirs issus des statistiques (modélisation) mais également de plus en plus de recherches en mathématiques appliquées. Elles visent notamment à développer des algorithmes pour analyser ces données (analyse de réseaux sociaux, apprentissage supervisé et semi-supervisé, etc.) Pour mieux comprendre ce processus de la science des données (*data science workflow*), on peut se référer à l'article de Dagiral & Parasie (2017), et au schéma qui récapitule les différentes opérations successives et itératives effectuées sur ces données.

ce long travail autour des données ? Ce sont ces différentes questions que nous avons explorées à partir du *hackathon* recherche République Numérique, qui constitue le terrain empirique de cette étude.

2- *Les préparatifs d'un hackathon* : retour d'expérience

Le *hackathon* recherche République Numérique s'est déroulé le 12 Décembre 2015 à Paris[16] et a été consacré à l'analyse des données de la consultation en ligne du projet de loi République Numérique (Septembre-Octobre 2015) . Cette loi, initiée par la secrétaire d'Etat Axelle Lemaire et son cabinet, a été pensée autour des trois éléments de la devise de la République française : *liberté* accrue pour la circulation des données et du savoir, *égalité* de droits pour les usagers du net et *fraternité* pour une société numérique ouverte à tous. Plusieurs thématiques étaient abordées par le projet de loi, comme celle de la protection de la vie privée, de l'ouverture des données[17] d'intérêt général (comprenant les données publiques) ou bien encore du libre accès (*open access*) aux résultats scientifiques issus de la recherche publique. La consultation en ligne s'est déroulée en amont des étapes habituelles d'adoption d'une loi car l'objectif était de recueillir l'avis des contributeurs et d'enrichir le projet de loi avant même sa présentation devant le Conseil d'État. Pendant deux semaines, cette consultation a provoqué une mobilisation forte (environ 21 000 contributeurs y ont participé[18]). Dans le cadre d'un projet de recherche doctoral mené par une des co-auteures sur l'une des thématiques abordées dans le projet de loi- le libre accès aux publications scientifiques[19]- ce moment particulier d'échanges constituait un terrain

[16] Ce hackathon s'est déroulé à La Paillasse « laboratoire de recherche ouverte et citoyen » http://lapaillasse.org/

[17] En français, on traduit *open data par ouverture des données* (Denis & Goeta, 2013)

[18] Suite à la consultation République Numérique, le site de la consultation République Numérique a fourni quelques chiffres concernant la participation en précisant : « Au total, ce sont 21 330 contributeurs qui ont voté près de 150 000 fois et déposé plus de 8500 arguments, amendements et propositions de nouveaux articles sur le site republique-numerique.fr. »

[19] Ce projet de recherche doctoral vise à étudier les dynamiques de reconfiguration du régime des savoirs avec le déploiement des technologies numériques en s'intéressant aux différentes significations données au terme d'*open* en sciences aujourd'hui. Une approche ethnographique a été menée auprès d'acteurs français impliqués dans les débats sur le libre accès aux publications scientifiques. La consultation République Numérique a été un des moments principaux de cette enquête. Cette

d'étude privilégié permettant d'accéder aux revendications de différentes parties prenantes et mieux comprendre les dynamiques de participation politique en ligne associée à cette phase consultative. Au vu des nombreuses contributions (plus de 8500 arguments[20] proposés sur le site), la consultation était aussi l'occasion de développer des méthodes numériques, pour pouvoir extraire, analyser et visualiser les actions des contributeurs sur le site (votes, commentaires, etc.). Dans cette optique, nous avons constitué un binôme (ingénieure de recherche/doctorante) afin de mettre en œuvre ces méthodes numériques nécessitant des compétences spécifiques en informatique. Mais au lieu de mener cette étude sur un des articles en particulier du projet de loi, nous avons proposé d'organiser cet événement dans une démarche de recherche ouverte et contributive en travaillant sur l'ensemble des articles de la consultation. Cette proposition a rapidement suscité un intérêt de la part de plusieurs acteurs[21] qui se sont impliqués par la suite en tant que partenaires à l'organisation de cet évènement et notamment à ces étapes préliminaires de constitution des jeux de données. En tant qu'instigatrices de ce *hackathon*, nous avons pris part à ses préparatifs et en proposons ici un retour d'expérience. Frappées par la richesse des échanges et des questionnements qui se sont déployés au fil de l'organisation de cette journée, ce *hackathon* s'est révélé être un cas d'étude propice pour poser un regard réflexif sur les pratiques de recherche et les processus méthodologiques mis en œuvre autour de ce type d'événements. Cette réflexivité s'est construite tout au long de l'organisation de l'événement jusqu'à la fin de la rédaction de cet article. Nos profils complémentaires (chercheure en SHS et ingénieure de recherche) ont enrichi ce dialogue et ont permis de croiser une compréhension technique fine de cette fabrique des données .

---

recherche s'est accompagnée d'une démarche réflexive méthodologique sur les pratiques de recherche numérique au sein des sciences sociales. Le hackathon a été un des éléments qui en a nourri l'analyse.

[20] Le terme « argument » est employé sur le site de la consultation pour désigner un commentaire publié (pour ou contre) sur un article de loi. Cela représente une option de participation sur le site de la consultation en plus du vote, de l'ajout de références ou bien encore de la proposition de nouvelles versions d'un article du projet de loi.

[21] L'intérêt porté à cet évènement était lié au projet de loi République Numérique et à la mobilisation importante qu'il a suscité avec plusieurs articles de loi vivement débattus par un ensemble de parties prenantes (enjeux économiques de l'ouverture des données, thématiques de la neutralité du web, entrée dans la loi d'un droit positif sur les communs). Face au grand nombre de contributions, le hackathon représentait un moyen pour différents acteurs de mieux comprendre les différents points de vue qui s'étaient exprimés et les dynamiques de mobilisation engendrées.

Revenons d'abord sur les origines de cette journée, afin de présenter les acteurs principaux qui se sont investis par la suite dans la préparation des données. Etalab, service ministériel dont la mission est de faciliter l'ouverture de données publiques, avait organisé un *Open Data Camp*[22] sur la vie quotidienne des données le 17 octobre 2015, soit 2 mois avant l'événement analysé ici. Un groupe de participants avait alors proposé d'explorer les données issues de la consultation République Numérique. Notre présence à ce *camp* nous a permis d'entrer en contact avec Etalab pour leur proposer d'organiser le *hackathon* recherche République Numérique (#hackRepNum) début novembre 2015. Notre rôle a tout d'abord consisté à trouver des partenaires pour réserver un espace de travail et financer quelques dépenses (repas et pause offerts aux participants). Nous avons donc soumis une proposition par *mails* à différents acteurs mobilisés autour de la consultation République Numérique, qui se sont vite montrés intéressés par un tel événement, et cela pour diverses raisons.

Des acteurs institutionnels (Etalab, le cabinet d'Axelle Lemaire à l'initiative de ce projet de loi mais aussi le Conseil National du Numérique) s'y sont montrés intéressés du fait de leur important investissement dans ces nouvelles logiques contributives, ouvertes et démocratiques. Cette consultation a ainsi fait l'objet d'une promotion importante fondée sur la mise en avant de l'originalité d'une telle initiative dans la fabrique de la loi et du développement d'une plateforme pour faciliter la participation citoyenne. Avant la consultation, le Conseil National du Numérique s'était attelé à la proposition d'un pré-projet de loi (Rapport Ambition Numérique), rédigé lui-même suite à une phase de six mois de consultation citoyenne et d'ateliers participatifs. La plate-forme de la consultation République Numérique en elle-même a fait l'objet d'un appel à projet auprès des acteurs impliqués dans les *civic tech*. Ces technologies civiques regroupent aujourd'hui un ensemble d'initiatives en lien avec le développement de plateformes numériques dont l'objectif est d'améliorer les processus démocratiques et la participation des citoyens (Kreiss, 2015). L'entreprise Cap Collectif, l'une de ces

[22] Nous ne reviendrons pas dans cet article sur ce format *Camp*, mais il s'agit d'un autre format court développé dans l'univers des start-ups et de l'innovation technologique. Inspiré des BarCamps, ces camps se caractérisent par leur organisation spécifique : on y tient des « non-conférences » où les échanges se font sans planning défini et où sont rédigés et partagés en direct le compte rendu des échanges. Les humanités numériques (Digital Humanities), se sont inspirées de ce format d'événement non conventionnel en développant les THATCamp pour « The Humanities and Technology Camp ». Ces formats ont même participé à l'émergence de cette nouvelle « transdiscipline » mêlant technologies et études sur les humanités. La tenue du premier THATCamp français a par exemple été l'occasion de la rédaction collaborative du Manifeste des Digital Humanities (Dacos, 2011).

« start-up civique », spécialisée dans le développement de plateforme de consultation en ligne, a été ainsi choisie aux côtés d'autres acteurs pour développer la plateforme (*Democracy OS*). L'un de ses responsables, présent le matin du *hackathon* aux côtés d'autres acteurs des *civic tech,* voyait ainsi cette journée comme un moyen de recueillir des retours pour améliorer ces plateformes contributives[23]. Cet événement représentait aussi un intérêt pour les collectifs impliqués dans les revendications autour des « communs » et de la libre circulation de l'information (mouvements du logiciels libres, de l'*open, etc.)*. Lors de la consultation, ces thématiques étaient au coeur des débats (utilisation de logiciels libres dans l'administration publique, reconnaissances des communs comme droit positif mais aussi conditions d'ouverture des données publiques[24]). Pour des chercheurs en SHS, cette consultation constituait un terrain d'étude pertinent pour comprendre les formes de mobilisation en ligne, les éléments saillants des débats ou bien encore les stratégies communicationnelles déployées par chaque participant pour faire valoir leur point de vue.

Après avoir reçu des premières réponses positives, la date de début décembre 2015 a été choisie. S'en est suivi un mois intense de préparation. Ces préparatifs ont d'abord porté sur l'organisation de la communication et des interventions des différents partenaires en début et fin de soirée (responsable de la consultation République Numérique, co-fondateur de la start-up Cap Collectif, responsable du Conseil National du Numérique et d'Etalab, fondateur de collectif ou d'association, etc.). Mais cette préparation a aussi consisté à prendre part aux nombreux échanges pour délivrer des jeux de données analysables pour les participants au *hackathon.* Même si nous n'avions pas employé le terme de *data sprint*, l'objectif, face au temps très court qui nous était imparti, était en effet de concentrer les participants sur

[23] Le site web de la consultation proposait une lecture facilitée de chaque article de la loi, présenté de façon claire et distincte (pour éviter le document brut et souvent illisible caractéristique aux textes de loi). Pour chaque article, quatre actions étaient possibles : le vote (pour, contre ou mitigé), l'ajout d'une modification à l'article initial (proposition d'un nouvel article ou un amendement), le commentaire (arguments pour ou contre sur l'article initial ou sur les modifications) et le partage de ressources supplémentaires.

[24] Pour citer quelques exemples, c'est le cas de l'article concernant le libre accès aux publications scientifiques qui discutait des modalités de partage et de diffusion des résultats de la recherche issus de fonds publics. Chercheurs, instituts de recherche, syndicats de l'édition ont pris part à cette consultation pour faire peser leur avis sur l'orientation précise de cet article. Une autre thématique qui a suscité de nombreux débats était celle de la définition et de la reconnaissance de ces communs informationnels comme droit positif.

l'analyse de données préalablement collectées. Quelques heures seulement se sont écoulées entre le choix des projets, la constitution des équipes en début de journée et la diffusion des premiers résultats en soirée, devant les divers partenaires et acteurs institutionnels.

Ces préparatifs engagent en fait une temporalité bien plus longue. Durant un mois, il s'est agi à la fois d'extraire les informations du site, de constituer des jeux de données et d'en assurer l'ouverture. Un travail important de « mise en données » qui révèle une ensemble d'enjeux socio-politiques, inclus dans les choix mêmes – et notamment les choix techniques – opérés par les acteurs impliqués.

3- Le travail de mise en données : des choix techniques révélateurs de choix politiques

Le terme de *datafication* est employé aujourd'hui pour désigner « la fabrication des matériaux à partir desquels sont produits statistiques et jugements de faits »(Bastin & Francony, 2016)[25]. Le *hackathon* nous a permis d'observer ce processus mais aussi d'en préciser quelques étapes. Dans le cadre de ces événements, en plus des étapes de sélection et d'extraction d'informations du site web pour en constituer des jeux de données, une autre étape essentielle consiste en leurs mises à disposition et leur ouverture éventuelle dans un format approprié. Chacune de ces étapes s'accompagne de différents choix techniques, révélateurs d'un ensemble d'enjeux socio-politiques associés à ces *data* et à leur réutilisation.

5-1 Acquisition des données : négociations, techniques d'extractions et hack

Différentes méthodes de collecte des données

L'étude d'un site web ou d'autres dispositifs numériques commence par une première étape : la collecte des informations jugées nécessaires à l'analyse. Dans le cas du site

[25] L'étude (Bastin & Francony, 2016) montre ainsi que chaque acteur en présence apporte sa propre interprétation de ce qu'est une donnée du web (chercheur, informaticien mais aussi plateforme et régulateur public) et de la manière dont il faut la produire.

de la consultation, ces informations pouvaient par exemple correspondre aux différentes activités possibles à réaliser sur la plateforme (vote, commentaires pour un article de loi et leurs contenus textuels) ou aux renseignements sur les participants disponibles en ligne via la page profil des inscrits. Cette collecte implique une acquisition de données en vue de les organiser en jeux de données exploitables. Les méthodes employées pour le *hackathon* résument bien les différentes possibilités offertes aujourd'hui par ces techniques que l'on regroupe souvent par le terme de « *text and data mining* », ou plus spécifiquement de « *web mining*» dans le cas de données exposées sur un site web.

La première consiste à demander aux concepteurs d'un site de fournir ces informations[26]. Dans le cas de la consultation République Numérique, ces informations ont été demandées par le cabinet ministériel avec l'aide d'Etalab à Cap Collectif en tant que prestataire de service. Mais d'autres moyens existent pour accéder à ces informations en fonction des compétences techniques des personnes mobilisées, ainsi que des outils mis à disposition. La deuxième méthode d'acquisition consiste à employer une API (*Application Programming Interface*). Cette interface de programmation, développée par le détenteur du site ou offerte par la plateforme elle-même comme mécanisme d'exposition des données, permet d'accéder aisément à certains éléments de la base de données internes du site. Concernant la consultation, Cap Collectif n'avait pas développé d'API spécifique à destination des développeurs. Les personnes désireuses d'extraire ces données se sont alors appuyées sur une troisième technique, plus « artisanale », qui consiste à « *scraper* [moissonner] » les données provenant d'un site web. Cette méthode est souvent qualifiée de « sauvage », étant effectuée par un développeur sans passer par une API. Dans le cas de la consultation, des ingénieurs et informaticiens avaient déjà débuté cette collecte, en particulier lors de l'*Open Data Camp*, mentionné précédemment. Le choix de l'une ou l'autre de ces méthodes conditionne les possibilités d'accès puis de constitution des données. Arrêtons-nous sur la dernière, dite de « *scraping* sauvage », car elle est associée à des formes d'action politique qui puisent dans la culture du hack et dont cette journée en a révélé l'importance.

[26] Pour rappel, un site web ou une application mobile repose sur une base de données où l'ensemble des informations (architecture du site, contenu des pages) et les opérations effectuées sur le site sont stockées. Ces informations sont disponibles par le concepteur du site.

Scraping sauvage : des choix techniques et politiques puisant dans la culture du « hack »

Cette méthode de *scraping* est qualifiée de « sauvage » en ce qu'elle est la plupart du temps, d'un point de vue juridique, considérée comme illégale : il s'agit d'exploiter des « œuvres » (sites web, applications, etc.) dont les données exposées sont protégées par le droit d'auteur ou par les droits des bases de données. L'emploi même de ces techniques implique également de répéter fréquemment des requêtes d'accès au serveur informatique, pouvant être assimilées à des attaques ou des intrusions dans le système informatique[27]. Cette méthode de *scraping* nécessite des compétences en informatique : elle consiste à développer un script d'extraction, autrement dit un algorithme, qui prend la forme d'une succession d'instructions à la « machine », écrit dans un langage de programmation spécifique (par exemple R, Python, etc.). Le développeur va s'appuyer sur le code html des pages pour découper les éléments qu'il souhaite obtenir (d'où le terme de *scraping*) et d'en automatiser l'extraction. Dans le cas de la consultation, cela pouvait être le nom d'un contributeur, l'identifiant d'un l'article de loi, la date de contribution, le nombre de vote pour un article, etc. Ces informations sont extraites en simulant le parcours d'un utilisateur, c'est-à-dire en créant un script qui mime les « clics et actions » d'un contributeur. On parle alors de « bot » (robot). Le développeur décide ainsi ce qu'il souhaite récupérer, mais aussi le point d'entrée qu'il juge le plus pertinent. Dans le cas de la plateforme de la consultation République Numérique, plusieurs points d'entrée ont été exploités : les pages utilisateurs, les pages des articles ou les pages historiques. Cette méthode offre une plus grande liberté que l'emploi d'une API, cette dernière, développée par un concepteur, exposant certaines informations plutôt que d'autres, et souvent avec un autre objectif que celui de les mettre à disposition. L'objectif principal est souvent d'obtenir des *analytics* sur l'utilisation du site (parcours de navigation sur le site, nombres de pages consultées, etc.). Ces API opèrent ce que nous appelons une

[27] Sauf si mention contraire, les données produites par le dispositif technique de la plateforme appartiennent au prestataire de service (la plateforme) ou au client final (le mandataire). Les licences spécifiques telles que les licences Creative Commons ou ODBL mentionnées sur les sites précisent lorsqu'elles sont présentes les conditions de partage et de réutilisation des informations disponibles sur un site.

première « mise en vue » des données par le propriétaire du site, orientant la collecte d'informations en mettant en valeur certaines informations (filtre de lecture), au détriment d'autres jugées comme moins importantes.

En proposant de sortir du « programme d'action » établi par ces API, la méthode de *scraping* peut ainsi jouer le rôle d'une réappropriation des informations et d'une plus grande flexibilité d'extraction. Jean-Christophe Plantin (Plantin, 2014) qualifie cette méthode de « polémologie poétique » (du grec *polemos*, guerre, et *logos*, discours), qui rejoint cette culture du « hack » associée à la revendication d'une liberté de l'information. Dans cette logique, le hack porte à la fois une dimension politique, la « libération » des données étant un moyen d'offrir une alternative à la lecture officielle. Une visée sociale et éthique y est également souvent associée, l'objectif étant de rendre ces *data* disponibles au plus grand nombre, à commencer par les personnes concernées par ces données (les contributeurs). Elle s'accompagne également, dans une visée de transparence, du partage de l'ensemble des informations permettant d'obtenir ces données.

Dans le cas du *hackathon* République Numérique, plusieurs personnes se sont attelées à cette tâche et s'inscrivent dans différentes dimensions du hack et de ses modalités d'actions. Nous avons nous-même travaillé à un script d'extraction de ces données en utilisant une partie du travail réalisé par des membres d'un collectif associatif, Regards Citoyens. Ce collectif se donne pour mission de proposer un accès simplifié au fonctionnement des institutions démocratiques à partir des informations publiques. Cette association mène ainsi un ensemble d'actions afin de rendre ces informations disponibles, mais également d'en faciliter la lecture par le biais de représentations graphiques ou d'interfaces. Regards Citoyens s'est également mobilisé lors de la consultation, au côté d'autres collectifs de liberté de l'information et des communs[28]. Certains membres du collectif ont travaillé avant même le *hackathon* sur l'extraction des données de la consultation, et ont mis à disposition les scripts sources, les données sources utilisées mais également les données

[28] Une vingtaine d'association dont Regards citoyens et d'autres partenaires de l'évènement se sont exprimés lors de la phase consultative pour soutenir les (biens) communs. Un site récapitulait l'ensemble des propositions soutenues par ces organisations http://soutenonslesbienscommuns.org/contributions/

retraitées[29]. Ce partage des scripts d'extraction, mais aussi des « données brutes » et retraitées, s'inscrit dans un mode d'organisation propre au développement informatique en tant qu'il permet d'améliorer le code, de faire gagner du temps à d'autres développeurs et d'en favoriser la reprise et la réutilisation. On y retrouve la dimension *sprint* des méthodologies agiles de développement logiciel abordé en début d'article. Mais ce partage se place également dans cette éthique du *hack* suivant laquelle la circulation de l'information se doit d'être traçable et reproductible.

Ces méthodes d'extraction dévoilent ainsi, derrière un ensemble de subtilités techniques, une dimension politique forte. La méthode de *scraping* sauvage illustre un ensemble de pratiques de collaboration, de traçabilité de l'information et d'accès aux sources (scripts, documentation du code, données) dans une visée politique qui puise dans une culture du hack. Mais pour le concepteur d'une plateforme, le choix de mettre à disposition une API et d'en présenter certaines informations consistent également en une maîtrise possible – consciente ou inconsciente – de la diffusion des informations et de leurs exploitations futures. L'utilisation de ces méthodes de *Text and Data mining* reste cependant légalement très peu encadrée. La loi République Numérique entendait statuer sur leurs contours en définissant les conditions d'utilisation de telles collectes, notamment dans le milieu de la recherche. Le manque de cadrage aidant, ces pratiques de collecte sont communément tolérées et souvent admises *de facto* dans les *hackathons* organisés par les institutions, elles-mêmes portées par l'injonction de faciliter l'ouverture, l'utilisation et la circulation de ces données. C'est aussi autour de cette mise à disposition des données et de leur «ouverture» que d'autres leviers politiques apparaissent.

5-2 Plus qu'une mise en donnée : les enjeux de de la mise à disposition et de l'ouverture

[29] Les informations sont disponibles sur cette page : https://regardscitoyens.github.io/contributions-PJLNum/web/.

La mise en ligne des données sous différents formats, via le développement de l'*open data*, fait l'objet de nouveaux enjeux concernant la fabrique des « données brutes »(Denis & Goëta, 2017). Leur mise à disposition, notamment sur un portail *open data*, nécessite que les données soient mises en forme, c'est-à-dire présentées de façon organisée dans un format de fichier qui pourra être ouvert par un logiciel approprié. On parle alors de « jeux de données » car ces données apparaissent sous un format standardisé et facilement consultable. Deux façons d'organiser les données sont souvent proposées suivant des standards informatiques : une vue tabulaire ou une vue en arborescence, dans le cas de données du web. Ces mises en forme, si anodines qu'elles paraissent, portent en elles-mêmes des types de « mises en vue » qui peuvent influencer l'interprétation des données. Mais ce processus peuvent être également influencé par l'anticipation des usages qu'il sera fait de ces données.

Le cas du jeu de données officiel mis en ligne quelques heures avant le début du *hackathon*, sous une forme anonymisée, est particulièrement représentatif du processus même de fabrique de ces données ouvertes. Dans le cadre de la loi République Numérique, Etalab a été chargé d'opérer la mise à disposition du jeu de données de la consultation sur le portail *open data* data.gouv.fr. Cette volonté du cabinet ministériel, client de la plateforme, de rendre publique les données de la consultation s'inscrit dans un discours sur l'ouverture et la transparence du gouvernement, dont le dispositif de consultation lui-même devait être l'un des reflets. L'organisation du *hackathon* a révélé l'intérêt de différentes communautés pour ces données, et a joué un rôle important dans leur mise à disposition en *open data*, peu de temps après cette consultation. Comme souligné précédemment, la constitution d'un jeu de données nécessite un ensemble d'étapes longues et chronophages. Ce processus peut-être encore plus long dans ce contexte institutionnel et officiel, en raison des différents intermédiaires qui y participent. Ici, les échanges impliquaient le cabinet ministériel, qui voyait un avantage à bénéficier d'une visibilité supplémentaire en termes de communication sur la consultation, Etalab, responsable de cette mission *open data*, et enfin le prestataire de service Cap Collectif. Etalab a été chargé de la négociation, mais aussi de la fabrication et de la mise en conformité des données. Après avoir demandé l'accès aux données du site, un temps supplémentaire a été nécessaire pour les mettre en forme ainsi que pour les dés-identifier afin de respecter

l'anonymat des participants. Ces opérations constituent également des mises en vue spécifiques de ces données, influençant l'interprétation possible qui en sera faite. Les données se présentaient dans un format tabulaire proche de la base de donnée proposée par le détenteur du site. Or, cette base de données suggère d'ores et déjà une organisation des informations afin d'obtenir un ensemble de mesures sur l'usage du site web par des internautes le consultant[30]. Cette orientation de lecture se retrouve ensuite dans la forme tabulaire du jeu de données mis à disposition.

L'anonymisation des données constitue un autre traitement effectué sur ces informations, qui consiste à supprimer des données personnelles – indicateurs socio-démographiques, de géolocalisation - permettant d'identifier une personne à partir de ces jeux de données. Elle représentait une nécessité juridique et éthique pour le gouvernement, mais aussi un enjeu politique pour d'autres acteurs. Si l'on s'attarde sur les autres données mises en ligne lors du *hackathon*, on constate en effet que les données fournies par Regards citoyens n'étaient pas anonymisées, ce qui s'apparente à un choix politique assumé. Celui-ci a été expliqué par les membres du collectif lors de la présentation de leur résultat à la fin du *hackathon*. Pour eux, ces informations étant publiques[31] lors de la consultation, il convenait de les laisser disponibles dans les jeux de données. Les cartographies réalisées à partir de ce jeu de données non anonymisé donnaient ainsi un accès facilité aux noms des auteurs auto-déclarés (selon leur profil sur la plateforme), mais offraient aussi une catégorisation par communauté, construite par Regards Citoyens. Ces cartographies suivaient entre autres l'objectif de mettre en lumière l'action de lobby d'acteurs spécifiques lors de la consultation. Cet exemple illustre l'influence même de l'utilisation d'un jeu de données plutôt qu'un autre sur l'analyse et l'interprétation qu'il pourra être fait des *data.*

[30] On emploie souvent la notion d'Analytics (Google Analytics, Facebook Analytics) pour désigner ces mesures d'usage d'un site web. Ces mesures contiennent par exemple le nombre de pages visitées, le temps passé sur chaque page, les origines géographiques des personnes ayant consulté le site ou bien encore les chemins de navigation les plus fréquents sur un site donné.

[31] Un utilisateur était libre d'employer un pseudonyme sur la plateforme en s'inscrivant.

Conclusion

En nous centrant sur l'étude du *hackathon* République Numérique, notre propos était de souligner des choix politiques militantes ou institutionnelles associées à la fabrique des données, qui opèrent en amont de ces évènements. Ce type d'événements consiste le plus souvent à exploiter (analyser, visualiser, etc.) des jeux de données préalablement constituées, soit proposées par les organisateurs, soit directement accessibles en ligne sur des portails publics de type *open data*. En revenant sur la mise en œuvre d'un tel événement et de sa préparation, c'est tout un pan de la fabrique de ces données, d'ordinaire invisibilisée ou normalisée, que nous avons tâché de révéler. De la collecte à la mise à disposition et à l'ouverture éventuelle de jeux de données, un ensemble de choix a été décrit, mettant en avant les dimensions politiques encapsulées dans ces pratiques informatiques, qui pour certaines puisent dans des formes de mobilisation et d'activisme politiques ancrées dans la culture du hack.

Certes, la nature hautement politique des informations contenues sur le site de la consultation République Numérique a probablement provoqué un coup de projecteur sur la dimension politique de cette fabrique des données. Si celle-ci peut être moins présente lors d'autres événements, nous défendons l'idée que toutes les données numériques analysées lors de ces formats courts portent une dimension politique, ne serait-ce que par leurs origines et par les conditions de leur mise à disposition par divers acteurs publics ou privées (données Facebook, Twitter, données *open data*).

Cet article invite donc les chercheurs qui organisent ou participent à de tels événements à porter un regard réflexif sur les étapes constitutives de la fabrication de ces jeux de données. Quelles sont les implications de l'utilisation d'une méthode ou d'une autre d'extraction des données ? Quelles influences jouent la présence d'une API ou d'un portail open data pour collecter ces données ? Quel cadre juridique faut-il construire pour ces recherches ? Et quelle position le chercheur doit-il tenir, pour l'utilisation de ces données, par rapport aux acteurs publics et privés ? Ces questions éclairent les enjeux sociaux, politiques et techniques dont le chercheur doit se saisir. Elles peuvent être associées à cette forme nécessaire d'« ouverture critique » dont

Serge Proulx nous amenait déjà à considérer, en 1984, dans son introduction à un numéro de *Sociologie et Sociétés* consacré à l'informatisation. Une attitude qui *« se veut lucide et critique en questionnant fondamentalement l'articulation entre d'une part, le développement de nouveaux usages sociaux des objets techniques et d'autre part, les nécessités et les besoins vitaux de nos pratiques sociales* » (Proulx, 1984, p.11). Suivant le développement de ces formats courts dont nous avons ici étudié une variante, adopter une telle posture nous semble plus que jamais d'actualité.

Bibliographie

BASTIN, G. et J-M FRANCONY (2016). « L'inscription, le masque et la donnée », *Revue d'anthropologie des connaissances*, vol. 10, n° 4, p. 505-530.

BIGOT, J-É et C. MABI (2017). « Une instrumentation numérique des sciences humaines et sociales », *Les Cahiers du numérique*, vol. 13, n°3, p.. 63-90.

BOULLIER, D (2016). *Sociologie du numérique*. Paris, Armand Colin.

BOWKER, G. C. et S. L STAR (2000). *Sorting things out: Classification and its consequences*. Cambridge, MIT press.

BRIATTE, F. et S. GOËTA (2014). « Les logiques politiques de l'ouverture des données de santé en France », *Statistique et Société*, vol. 2, n°2, p. 49-55.

BRISCOE, G. et C. MULLIGAN (2014). « *Digital innovation: The hackathon phenomenon* »,. Présenté à The Sustainable Society Network+ 2014, Queen Mary University of London.

BROCA, S. (2013). *Utopie du logiciel libre*. Neuvy-en-Champagne, Le Passager Clandestin.

CARDON, D. et F. GRANJON (2013). *Médiactivistes*. Paris, Sciences po, les presses.

COLEMAN, G. (2011). « Hacker Politics and Publics », *Public Culture*, vol. 23, N°3 65, p. 511-516.

COLEMAN, G. (2012). « Code Is Speech: Legal Tinkering, Expertise, and Protest among Free and Open Source Software Developers », *Cultural Anthropology*, vol. 24 n°3, p. 420-454.

COLEMAN, G. (2014). « Hacker », in RYAN, M-L., L. EMERSON et B.J. ROBERTSON (dir.) *The Johns Hopkins Encyclopedia of Digital Textuality*, Baltimore, JHU Press, p. 245-248

DACOS, M. (2011). « Manifeste des Digital humanities », *Carnet de recherche THATCamp Paris*.

DAGIRAL, É. et S. PARASIE (2017). « La « science des données » à la conquête des

mondes sociaux : ce que le « Big Data » doit aux épistémologies locales », in Menger, P-Ml et S. PAYE (dir.), *Big data et traçabilité numérique : Les sciences sociales face à la quantification massive des individus*. Paris, Collège de France, p. 85-104.

DAGIRAL, É et A. PEERBAYE (2012). « Les mains dans les bases de données », *Revue d'anthropologie des connaissances*, vol. 61, n°1, p. 191-216.

DENIS, J. et S. GOËTA (2014). *Exploration, Extraction and 'Rawification'. The Shaping of Transparency in the Back Rooms of Open Data* (SSRN Scholarly Paper No. ID 2403069). Rochester, NY: Social Science Research Network.

DENIS, J. et S. GOËTA (2017). « La fabrique des données brutes : Le travail en coulisses de l'open data », in MABI, C., J-C PLANTIN et L. MONNOYER-SMITH (dir.), *Ouvrir, partager, réutiliser : Regards critiques sur les données numériques*. Paris, Éditions de la Maison des sciences de l'homme.

DESROSIÈRES, A. (2008). *Pour une sociologie historique de la quantification*. Paris, Presses de l'École des Mines.

FLICHY, P. (2013). « Rendre visible l'information », *Réseaux*, n° 178-179 (2), p. 55-89.

GITELMAN, L. (2013). *Raw data is an oxymoron*. Cambridge, MIT Press.

GOETA, S. (2016). *Instaurer des données, instaurer des publics : une enquête sociologique dans les coulisses de l'open data* (doctorat). Paris,Télécom ParisTech.

GRANJON, F. (2015). « Du pragmatisme et des technologies numériques, On pragmatism and digital technologies », *Hermès, La Revue*, vol. 73, p. 219-224.

GRANJON, F. (2017). *Mobilisations numériques: Politiques du conflit et technologies médiatiques*. Paris, Presses des Mines.

GRUSON-DANIEL, C. et C. MABI (2017). « AAC : « Formes et mouvements politiques à l'ère numérique » », *RESET*, n°7.

HACKEBOLA WITH DATA (2015). « HackEbola with Data: On the Hackathon Format for Timely Data Analysis », *Procedia Engineering*, vol.107, p. 377-386.

HEATON, L. et F. MILLERAND (2013). « La mise en base de données de matériaux de recherche en botanique et en écologie », *Revue d'anthropologie des connaissances*, vol. 7, n° 4, p. 885-913.

JATON, F. et D. VINCK (2016). « Processus frictionnels de mises en bases de données », *Revue d'anthropologie des connaissances*, vol. 10, n° 4, p. 489-504.

KREISS, D. (2015). « The Problem of Citizens: E-Democracy for Actually Existing Democracy », *Social Media + Society*, vol.1, n°2, p. 1-11.

LOVELUCK, B. (2015). *Réseaux, libertés et contrôle: une généalogie politique d'internet*. Paris, Armand Colin.

MABI, C.(2013). « Inclusion des publics et matérialité des dispositifs participatifs », *Participations*, n°7, p. 201-213.

MABI, C. et J-C PLANTIN (2017). « Introduction. Lorsque la recherche en sciences humaines et sociales se penche sur les données numériques », in MABI, C., J-C PLANTIN et L. MONNOYER-SMITH (dir.), *Ouvrir, partager, réutiliser : Regards critiques sur les données numériques*. Paris, Éditions de la Maison des sciences de l'homme.

MAUZ, I. et C. GRANJOU (2011). « Rendre visibles les « travailleurs invisibles » ? », *Terrains & travaux*, n° 18, p. 121-139.

MENGER, P-M et S. PAYE (dir.) (2017). *Big data et traçabilité numérique : Les sciences sociales face à la quantification massive des individus*. Paris, Collège de France.

MILLERAND, F. (2011). « Le partage des données scientifiques à l'ère de l'e-science : l'instrumentation des pratiques au sein d'un collectif multidisciplinaire », vol. 18, n°1, p. 215-237.

PEUGEOT, V. (2014). « Les Communs, une brèche politique à l'heure du numérique », in MARYSE, C. et J-M, NOYER (dir.), *Les débats du numérique*. Paris, Presses des Mines, p. 77-98.

PLANTIN, J-C (2014). *La cartographie numérique*. London, ISTE editions.

PLANTIN, J-C et L. MONNOYER-SMITH (2013). « Ouvrir la boîte à outils de la recherche numérique », *tic&société*, vol. 7, n° 2.

PROULX, S. (1984). « Présentation : L'informatisation : mutation technique, changement de société? », *Sociologie et sociétés*, vol. 16, n°1, p. 3-12.

SCHROCK, A. R. (2016). « Civic hacking as data activism and advocacy: A history from publicity to open government data », *New Media & Society*, vol. 18, n°4, p. 581-599.

STAR, S. L. et J. R. GRIESEMER (1989). « Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39 », *Social Studies of Science*, vol. 19, n°3, p.. 387-420.

SUTHERLAND, J. et K. SCHWABER (2013). *Le guide Scrum - Le guide définitif de Scrum : les règles du jeu*.

TURNER, F. (2012). *Aux sources de l'utopie numérique : De la contre-culture à la cyberculture, Stewart Brand, un homme d'influence*. Caen, C&F Edition.

VENTURINI, T., A. MUNK et A. MEUNIER (2016). « Data-Sprinting: a Public Approach to Digital Research », in *Interdisciplinary Research Methods*. LURY C., P. CLOUGH et MICHAEL, M., FENSHAM, R., S. LAMMES, LAST, A. et E. UPRICHARD (dir.).

ZUKIN, S. et M. PAPADANTONAKIS (2017). « Hackathons as Co-optation Ritual: Socializing Workers and Institutionalizing Innovation in the "New" Economy », in *Precarious Work*. Bingley, Emerald Publishing Limited, p. 157–181.