Health Data Hub
Pourquoi tant de haine ?

Peu de projets français d'IA auront suscité autant de controverses. La récente plate-forme de données et de services cloud pour la recherche en santé s'est fait de nombreux ennemis. En cause : le choix de Microsoft comme prestataire cloud. Une polémique qui fait de l'ombre à un projet pourtant très prometteur pour les data sciences et le secteur français de l'e-santé. Mais ce choix n'est pas irréversible.

Une levée de boucliers des promoteurs du logiciel libre, des tweets incendiaires d'Octave Klaba, une nécessaire prise de parole de Cédric O, pour calmer le jeu… Le projet Health Data Hub a cristallisé le débat français autour de la souveraineté numérique. La cause de la polémique n'est pas son objectif, à savoir : proposer une plate-forme des données de santé et de services cloud, aux chercheurs et aux entreprises, pour développer de nouvelles solutions d'e-santé, basées sur l'IA. Ce qui bloque : c'est la méthode. Sans appel d'offres, Microsoft s'est vu confier la fourniture de l'infrastructure et des principaux services cloud. Au grand dam des acteurs français du secteur et des défenseurs du logiciel libre, qui voient le géant américain se positionner au cœur d'un projet hexagonal aux enjeux considérables, basé sur des données ultra-sensibles : celles de la santé des Français.

Pour les uns, ce choix révélerait un manque de maturité de l'offre cloud en Europe, encore incapable de rivaliser avec celle des géants américains. Pour d'autres, c'est un nouvel exemple d'une certaine politique publique française, accordant plus facilement sa confiance à de grands groupes, même étrangers, qu'à des PME, pourtant hexagonales.

Stéphanie Combes, directrice du projet Health Data Hub : « Nous disposons en France de nombreuses données de santé, parmi les plus riches au monde, mais elles demeurent relativement sous-exploitées. »

Pour comprendre l'*« affaire Health Data Hub »*, il convient de cerner le contexte et la nature même du projet. Pourquoi a-t-il été lancé ? En quoi consiste exactement la plate-forme et comment fonctionne-t-elle ? Et, bien entendu, quelles données y sont hébergées et comment sont-elles protégées ?

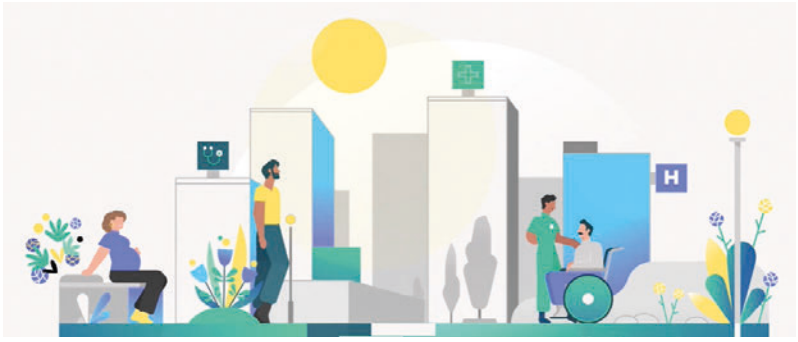
L'origine du projet est le rapport sur l'Intelligence artificielle du mathématicien et désormais homme politique Cédric Villani. Rendu au gouvernement en mars 2018, le document préconisait de *« créer une plate-forme d'accès et de mutualisation des données pertinentes pour la recherche et l'innovation en santé, regroupant dans un premier temps les données médico-administratives, puis les données génomiques, cliniques, hospitalières… »*

C'est à partir de cette recommandation qu'a été lancé le projet Health Data Hub en juin 2018. La situation de départ était la suivante : *« Nous disposons en France de nombreuses données de santé, parmi les plus riches au monde, mais elles demeurent relativement sous-exploitées, particulièrement par*

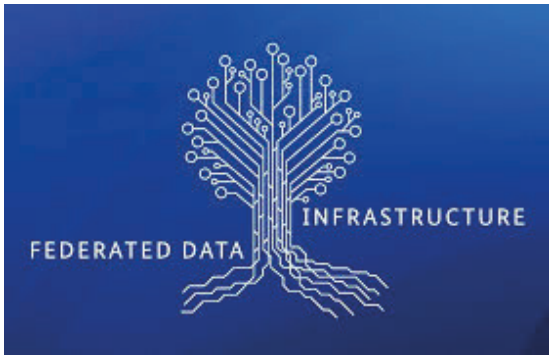
des approches d'IA et de data science. Pourquoi ? Tout d'abord car elles sont réparties sur un grand nombre de systèmes hétérogènes, avec des bases de données très cloisonnées. Ensuite, car ces systèmes ne disposent pas toujours des technologies nécessaires à l'exploitation de ces données avec des approches de data science. Ces technologies sont pourtant nécessaires pour la réalisation des traitements de gros volumes de données en Python, en R, en Spark… avec des capacités de calcul et de stockage importants », explique Stéphanie Combes, directrice du projet Health Data Hub.

L'objectif est donc de rassembler un maximum de données de santé sur une seule plate-forme, et d'y intégrer de la puissance de calcul (CPU et GPU), ainsi que des services cloud pour faire tourner des réseaux de neurones et autres algorithmes d'IA. Point important : il ne s'agit pas de centraliser toutes les données de santé françaises. Le Health Data Hub ne collecte que des copies de ces données, à des fin de data science, et non de suivi médical. *« Il y a souvent une confusion, mais le Health Data Hub n'est pas une plate-forme de production pour les professionnels de santé. Il s'agit d'une plate-forme de R&D destinée à de la recherche et uniquement à la recherche »*, souligne Stéphane Messika, CEO de Kynapse by open. Ce cabinet de conseil en stratégie digitale, filiale de l'intégrateur Open, a assuré la maîtrise d'œuvre du projet. Le Health Data Hub s'adresse donc aux acteurs de la recherche en santé publique (instituts, universités, laboratoires…) mais aussi aux « porteurs de projets » du secteur privé, à commencer par les start-up. *« L'intérêt de la plate-forme est d'offrir à la fois un environnement de services et des données. C'est la combinaison de ces deux éléments qui est source de valeur »*, indique le docteur Arnaud Rosier, fondateur de la jeune pousse Implicity, à l'origine du premier projet exploitant le Health Data Hub (lire encadré p. 12). L'accès à la plate-forme s'effectue via un espace dédié, totalement cloisonné où les utilisateurs peuvent exploiter les données liées à leur projet. *« Le plus compliqué a été de créer ces espaces sécurisés, élastiques et perméables avec le respect des normes de protection du système de santé français »*, poursuit Stéphane Messika. Avec ce système d'espaces cloisonnés, les données de santé ne sortent pas de la plate-forme et les porteurs de projets n'ont accès qu'aux datas qui les concernent.

Pour l'instant l'accès à la plate-forme n'est pas payant. Mais le Health Data Hub réfléchit à une tarification de son offre de services. *« Pour la recherche publique, la plate-forme resterait gratuite, mais pour le secteur privé, certains services pourraient être tarifés. Cela nous permettrait de redistribuer les revenus aux acteurs à l'origine de la collecte des données »*, confie Stéphanie Combes. Le business model du Health Data Hub est donc aujourd'hui basé quasi-uniquement sur de la subvention publique. Le budget s'élève à près de 80 millions d'euros sur quatre ans, dont 36 millions proviennent de l'appel à projets du fonds pour la transformation de l'action publique (FTAP) et 40 millions de l'Ondam (Objectif national de



Gaia-X : la marque du Cloud souverain européen



En juin 2020, vingt-deux acteurs du Cloud français et allemand ont lancé le projet Gaia-X, avec comme objectif de développer des offres cloud capables de concurrencer les Microsoft Azure et autres AWS. Il ne s'agit pas d'un vaste Cloud européen, basé sur des data centres interconnectés, mais d'une « marque », intégrant des critères que se sont engagés à respecter les membres fondateurs. *« C'est une marque de confiance d'un Cloud souverain européen, regroupant des offres avec des engagements autour de la portabilité, de l'interopérabilité, de la protection des données… »*, résume David Chassan de 3DS Outscale. À la différence du Cloud souverain français, Andromède, qui a échoué : *« Les membres fondateurs de Gaia-X ne sont pas que des cloud provider mais aussi des clients, comme EDF, Amadeus ou Safran. Cela nous garantit des commandes. »* Les premières offres Gaia-X sont attendues pour la fin 2020 - début 2021.

dépenses d'assurance maladie). Des partenaires privés ont également des contributions à hauteur de quelques dizaines de milliers d'euros. Mais elles restent donc à la marge.

Des données « pseudonymisées »

Les données stockées sur la plate-forme sont aujourd'hui principalement issues du Système national des données de santé (SNDS), qui regroupe l'ensemble des données de santé associées à un remboursement de l'Assurance maladie. Il s'agit des informations collectées à l'occasion d'une prise en charge à l'hôpital, d'une visite chez le médecin ou d'une participation à une cohorte de recherche. On y retrouve notamment la pathologie du patient, son traitement, les éventuels actes médicaux, etc. D'autres bases de données devraient être prochainement ajoutées, comme celles du Samu ou des pharmacies. *« Nous n'avons pas vocation à collecter toutes les données de santé du pays, mais seulement celles qui sont intéressantes pour la recherche »*, précise la direction du Health Data Hub.

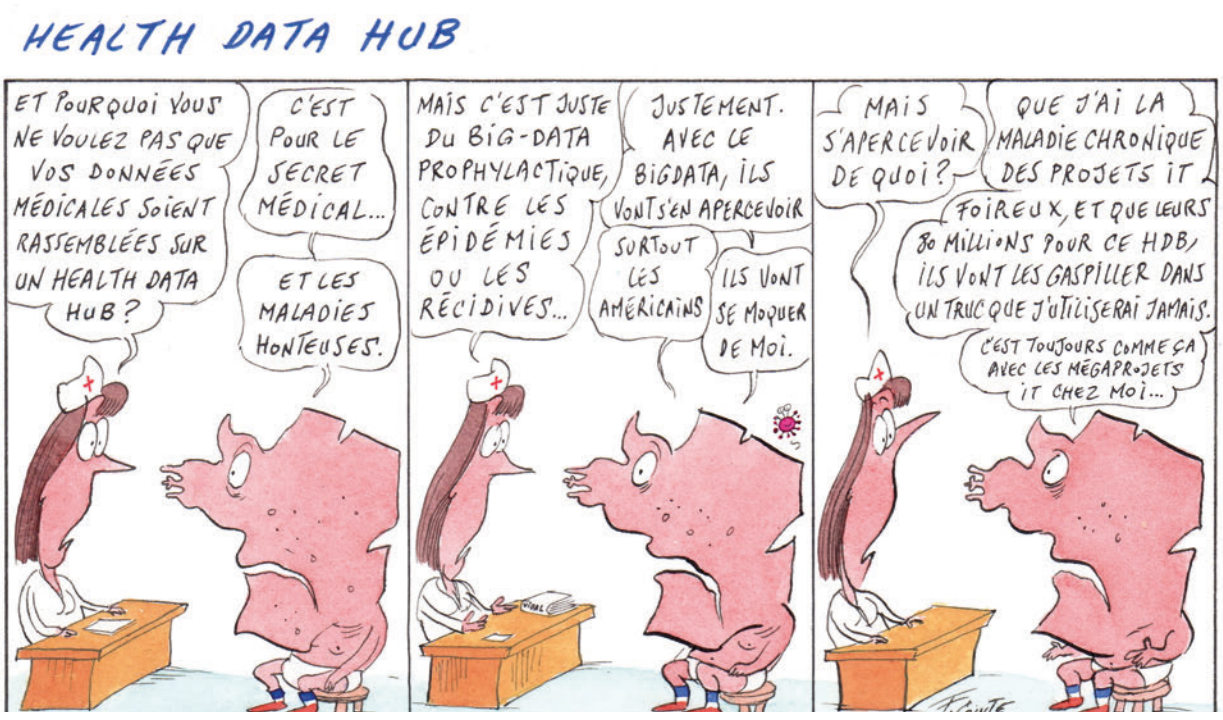
Ces données ne sont pas anonymisées mais pseudonymisées. La différence est subtile. Avec une donnée anonymisée, tous les éléments d'identidication sont détruits afin qu'il n'y ait aucune réversibilité. Avec des données pseudonymisées, seules les principales informations d'identification sont effacées (nom, prénom, date et lieu de naissance, adresse…). Il demeure donc des éléments spécifiques à l'individu. Par recoupements, il serait théoriquement possible de retrouver l'identité liée à des données pseudonymisées, comme l'ont démontré plusieurs études scientifiques. Mais ce type d'opération reste complexe. Le Health Data Hub explique avoir besoin de données pseudonymisées car la data science requiert l'analyse de « trajectoires individuelles », avec des successions d'événements (hospitalisation, complication, amélioration de l'état clinique, traitements pris…) ce que les données anonymisées ne permettent pas.

« Microsoft proposait les meilleurs outils managés et était le seul à être certifié HDS, notamment au niveau des GPU »

Stéphane Messika,
CEO de Kynapse by open

Un projet mené au pas de charge, d'où le choix de Microsoft

La première version de la plate-forme est opérationnelle depuis avril 2020. Le choix du prestataire a été pris début 2019 et la création du groupement d'intérêt public, gérant la plate-forme, date de novembre dernier. Pour un projet de cette envergure, le calendrier a été plutôt serré. Ce serait d'ailleurs pour



Cloud Act : la « loi du Far West »

Le *« Clarifying Lawful Overseas Use of Data Act »* ou Cloud Act est une loi fédérale américaine adoptée en 2018 sous l'administration Trump. Elle permet aux autorités américaines d'accéder à des données, dont des données personnelles, dans le cadre de leur procédure, en en faisant la demande auprès de fournisseurs de services, notamment cloud. Ses promoteurs y voient un cadre juridique pour une pratique de toute façon existante ; ses détracteurs, une porte ouverte au non-respect de la vie privée ou à l'espionnage industriel. Car le Cloud Act peut être invoqué sans aucune transparence sur la collecte et l'exploitation de données. Un juge n'est pas tenu d'informer les personnes concernées, ni de fournir des informations sur l'hébergement, l'utilisation et la sécurisation des données. *« Le Cloud Act pose de nombreux problèmes. Il entre en conflit avec le RGPD et ne permet aucun recours, puisque son exploitation est totalement opaque »*, déplore Alexandra Iteanu, avocate spécialisée dans l'IT. Comment être certain qu'un cloud provider, même français, n'est pas soumis à cette législation ? *« Il ne faut pas avoir son siège aux États-Unis, mais pas non plus de filiales sur place qui pourraient servir à capter des données en dehors du sol américain. Un acteur français qui possède une ou plusieurs entités aux États-Unis doit donc garantir qu'elles sont totalement cloisonnées techniquement et juridiquement. »* Il n'existe pas, à ce jour, de cas public d'exploitation du Cloud Act. *« Mais il y a des chances que le texte ait été utilisé. Nous n'avons cependant aucun moyen de le savoir »*, conclut Alexandra Iteanu.

Alexandra Iteanu, avocate spécialisée dans l'IT.

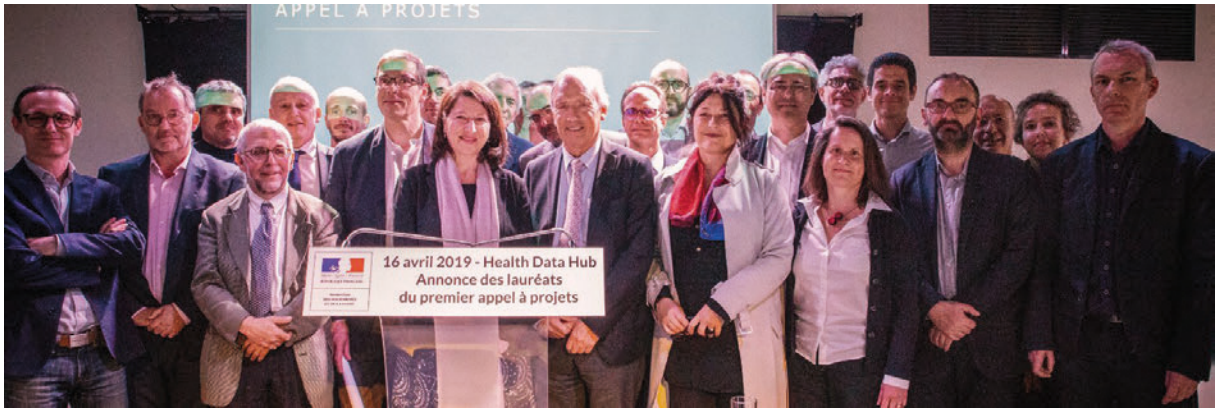
gagner du temps que le choix s'est porté sur Microsoft. Car, d'après la direction du Health Data Hub, seul Microsoft était prêt à l'époque. *« C'était le seul choix possible, au regard de nos attentes techniques et juridiques »*, indique Stéphanie Combes. D'où le fait qu'aucun appel d'offres n'ait été lancé, car seule la firme de Redmond aurait rempli tous les critères. *« Techniquement, Microsoft proposait les meilleurs outils managés et était le seul à être certifié HDS, ou Hébergeurs de données de santé, notamment au niveau des GPU »*, rappelle pour sa part Stéphane Messika.

Un point de vue que ne partage pas le CNLL (Conseil National du Logiciel Libre) qui a attaqué la Health Data Hub devant le Conseil d'État, début juin, avec une quinzaine d'autres organisations et personnalités, dont le collectif InterHop, Jean-Paul Smets chez Nexedi ou le médecin Didier Sicard : *« Pourquoi ne pas avoir attendu quelques mois afin qu'un ou plusieurs acteurs français puissent se positionner ? Microsoft était un choix facile, avec des solutions sur étagère, mais il est soumis au Cloud Act et cela pose tout de même un grave problème de souveraineté des données »*, estime Pierre Baudracco, co-président du CNLL et CEO de la société BlueMind. Rappelons que le Cloud Act permet à une juridiction américaine de collecter des données personnelles sur des suspects, sans aucune transparence sur l'exploitation de ces données (lire encadré). Autre problématique soulevée par les requérants, le projet n'intègre pas que la fourniture d'une plate-forme cloud *« mais aussi une cinquantaine de services managés sur Azure, dont on ne connaît pas précisément la nature et qui posent la question de la réversibilité de la plate-forme. Comment changer de prestataire si celle-ci a été construite sur des technologies Microsoft ? »*, pointe pour sa part Stéphane Fermingier, autre co-président du CNLL et CEO d'Abilian. Pour Jean-Paul Smets, président de Nexidi, des alternatives étaient possibles : *« Nous leur avons écrit pour leur dire que des technologies open-source étaient disponibles afin de construire la plate-forme. Et nous aurions pu répondre à un appel d'offres en nous rapprochant d'un Cloud provider. Mais nous n'avons eu aucune réponse. Le pire, c'est que toutes les technologies du type de celles utilisées sur Azure viennent de Français, comme Scikit-learn, une bibliothèque libre d'IA qui a été développée par l'Inria. C'est une négation des compétences françaises. »*

Le choix de Microsoft a également fait sortir de ses gonds Octave Klaba, le dirigeant d'OVH : *« C'est la peur de faire confiance aux acteurs français de l'écosystème qui motive ce type de décision. La solution existe toujours. Le lobbying de la religion Microsoft arrive à faire croire le contraire. C'est un combat. On va continuer et un jour on gagnera »*, a-t-il déclaré sur Twitter. Le 19 juin, le Conseil d'État a rendu une décision mitigée, qui est loin

« Le projet intègre une cinquantaine de services managés sur Azure, dont on ne connaît pas précisément la nature »

Stéphane Fermigier, co-président du CNLL et CEO d'Abilian



En avril 2019, Agnès Buzyn, ministre de la Santé, dévoilait les 10 lauréats du premier appel à projets du Health Data Hub.

d'avoir calmé les esprits. En résumé, la plus haute juridiction administrative française a estimé que le choix de Microsoft ne présentait pas de risques pour la protection des données privées. Il a cependant demandé au Health Data Hub de fournir de plus amples informations à la Cnil sur ses systèmes de pseudonymisation et de protection des données. *« Cette décision conforte le fait que la plate-forme technologique ne constitue pas un risque pour la vie privée des personnes »*, se félicite Stéphanie Combes. *« C'est une victoire car plusieurs contre-vérités, portées par le Health Data Hub, ont été mises en lumière »*, souligne pour sa part Stéphane Fermigier. *« Par exemple : nous avons appris que les données ne sont pas hébergées en France mais aux Pays-Bas. Par ailleurs, Microsoft peut avoir accès aux clés de chiffrement. »*
Dans l'autre camp, on rétorque vouloir une localisation des données en Europe. *« Et c'est le cas ! »*, indique Stéphane Messika. Concernant les clés de chiffrement, *« elles sont générées via un HSM, ou Hardware Security Module, externe et transmises à un HSM interne sur la plate-forme Azure de Microsoft. Le déchiffrement des clés est automatisé, il n'y a donc aucune intervention d'un administrateur de Microsoft dans cette opération. Et dans son contrat, Microsoft a l'interdiction d'utiliser ces clés »*, indique la direction du Health Data Hub. Sur les risques du Cloud Act, c'est une loi très spécifique qui permet de *« réquisitionner des données nominatives, à des fins d'enquête, et pas une base de données complète. Or, nous n'avons pas les données nominatives des patients. Elles arrivent déjà pseudonymisées dans la plate-forme »*. Enfin, sur la nature des cinquante services cloud associés à l'offre d'hébergement Azure, ils couvrent notamment *« la virtualisation de machines, la structuration des espaces de stockage, le déclenchement d'événements sur la plate-forme ou le transport des messages »*, indique la Health Data Hub. Environ quarante de ces services sont fournis par Microsoft et dix sont externalisés. *« Nous travaillons avec une dizaine de fournisseurs pour ces services, dont Wallix, éditeur français de logiciels de sécurité informatique, ou CDC Arkhinéo, spécialiste de l'archivage et de la conservation à long terme des données électroniques »*.

Implicity : première start-up à exploiter le Health Data Hub

Le projet « Hydro », porté par la jeune pousse Implicity, a reçu l'autorisation de la Cnil fin mai pour exploiter les données du Health Data Hub. Son objectif : développer un algorithme d'IA capable de prédire les crises d'insuffisance cardiaque pour les patients porteurs de prothèses cardiaques connectées (pacemakers et défibrillateurs). *« Ces équipements collectent des informations pertinentes, comme la fréquence cardiaque, la fréquence respiratoire et même la présence d'eau dans les poumons. L'algorithme pourra traiter ces données afin d'anticiper une crise environ 30 jours à l'avance. En modifiant en conséquence le traitement du patient, la crise sera évitée et il n'y aura pas d'hospitalisation »*, explique le docteur Arnaud Rosier, fondateur d'Implicity. Il rappelle que l'insuffisance cardiaque est la pathologie chronique la plus coûteuse en France. Elle représente la première cause d'hospitalisation des plus de 65 ans et 10 % des coûts globaux d'hospitalisation, soit 1,5 milliard d'euros annuels. *« Nous allons entrainer l'algorithme sur la plate-forme en l'alimentant avec les données d'hospitalisation que nous allons croiser avec celles de 20 000 patients télésuivis. »* Les travaux de R & D débutent en juillet. La start-up espère proposer une première version de son algorithme dans les six mois.

Docteur Arnaud Rosier, fondateur d'Implicity.

Dans ces conditions comment est-il possible de changer de prestataire ? *« Le choix de Microsoft a permis à la plate-forme une mise en production rapide, en moins de douze mois. Le moment venu, il sera possible de changer de prestataire, car la plate-forme technologique est conçue pour être réversible. En effet, nous utilisons la technologie Terraform pour scripter au maximum les travaux d'intégration et de déploiement dans une logique* Infrastructure As Code. *Le contrat passé avec Microsoft ne prévoit aucune clause d'engagement ou d'exclusivité »*, assure Stéphanie Combes.

Vers un prochain appel d'offres ?

Face à la polémique, Cédric O, secrétaire d'État au Numérique a déclaré qu'il *« serait normal que, dans les mois à venir, nous puissions lancer un appel d'offres »*. Un avis partagé par Guillaume Poupard, DG de l'Anssi, dont l'agence a participé au projet : *« Dans une phase de prototypage, le choix d'une solution facile d'emploi a été privilégié. Nous sommes maintenant dans une phase opérationnelle, et le fait de revenir sur une solution européenne idéalement qualifiée par l'Anssi et non soumise à des lois extra-territoriales européennes serait de bon goût. »* Plusieurs acteurs sont déjà sur les rangs, dont OVH qui a rappelé disposer aujourd'hui de la certification HDS et de celle de l'ANSSI. En lice également : Scaleway, le cloud d'Iliad. *« Le Health Data Hub peut même et d'ores et déjà travailler avec plusieurs fournisseurs de Cloud, simultanément, dont Scaleway et Microsoft ! Le choix n'est pas binaire »*, confie Yann Lechelle, son directeur général. Autre acteur positionné : 3DS Outscale (Dassault Systèmes) : *« Ce qui était vrai il y a un an, ne l'est plus. Aujourd'hui, nous sommes en mesure de répondre. Nous avons obtenu les certifications HDS et ANSSI fin 2019 »*, souligne son directeur de la stratégie, David Chassan. Quant à Orange : *« Au cas où l'État déciderait de lancer un appel d'offres pour le Health Data Hub, dans les prochains mois, Orange Business Services est en capacité de répondre aux exigences fonctionnelles, techniques et de sécurité connues à date sur les services cloud »*, confie Éric Pieuchot, directeur d'Orange Healthcare. Outre leurs diverses certifications, tous ces acteurs rappellent être membres fondateurs du projet de Cloud européen, Gaia-X (lire notre encadré). Microsoft aurait donc de la concurrence en cas d'appel d'offres.

Une trentaine de projets attendus en 2021

Malgré ce démarrage sous pression, le projet Health Data Hub entend poursuivre son développement, avec un rythme soutenu. *« C'est un projet de santé publique, mais il est aussi question de donner un avantage compétitif aux acteurs français de la santé. Or, dans le domaine des applications numériques de santé, qui évoluent très vite avec l'IA, les États-Unis et la Chine sont déjà très avancés. Il faut donc aller vite. C'est une question de souveraineté numérique »*, souligne Stéphanie Combes.

D'ici à 2021, la plate-forme devrait accueillir une trentaine de projets. Une dizaine a déjà été retenue par le Health Data Hub sur plus de 180 candidatures. Ils portent notamment sur l'évaluation et l'amélioration des parcours de soins après un infarctus du myocarde, la prédiction des trajectoires individuelles des patients parkinsoniens ou encore la quantification de la proportion de patients touchés par un effet médicamenteux indésirable. En cet été 2020, un seul a obtenu l'autorisation de la Cnil, celui d'Implicity (lire encadré).

L'interface devrait également prochainement évoluer pour gagner en ergonomie. *« Nous sommes pressés d'avoir des retours utilisateurs pour faire évoluer la plate-forme »*, indique Stéphane Messika. Le Health Data Hub prévoit aussi d'étoffer ses effectifs, aujourd'hui composés de 35 collaborateurs internes et d'une quinzaine de prestataires. Environ 25 recrutements sont prévus dans les six prochains mois, avec comme objectif d'atteindre 70 collaborateurs en 2021. Le Health Data Hub recrute notamment des juristes, des développeurs, des ingénieurs infrastructure cloud, des administrateurs système et des chefs de projets.

Mais l'enjeu principal pour la plate-forme est surtout d'étoffer son catalogue de données. *« D'ici à 2022, nous espérons proposer un catalogue suffisamment large pour être attractif vis-à-vis de l'écosystème de la recherche et de l'innovation »*, confie Stéphanie Combes. Elle évoque également sa participation à l'action conjointe de la Commission européenne pour la construction d'un European Data Space qui doit fédérer tous les Health Data Hub d'Europe. *« Le projet Findata, en Finlande, présente beaucoup de similarités avec le nôtre »*, conclut-elle. D'ici moins de 10 ans, un Health Data Hub européen pourrait ainsi voir le jour. Un projet qui ne manquera pas d'alimenter de nouveaux débats sur la souveraineté numérique et le respect des données personnelles ! ✖

Christophe Guillemin

Grâce à son infrastructure certifiée HDS et ANSSI, 3DS Outscale est un des candidats potentiels au futur appel d'offres du Health Data Hub.