



HAL
open science

Covid19 et Science ouverte, premiers reculs

Ghislaine Chartron

► **To cite this version:**

| Ghislaine Chartron. Covid19 et Science ouverte, premiers reculs. 2021. hal-03347094

HAL Id: hal-03347094

<https://hal.science/hal-03347094>

Preprint submitted on 16 Sep 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Covid19 et Science ouverte, premiers reculs

(A paraître, Editions EMS, Management & Société, janvier 2022)

Ghislaine CHARTRON, Professeur du CNAM, Chaire d'ingénierie documentaire

ghislaine.chartron@lecnam.net

Table des matières

Introduction.....	2
1. Explosion des partages de connaissances sur la Covid19.....	3
1.1 Impulsions politiques	3
1.2 Inflation et diversité des informations	4
1.2.1 Les données virales factuelles.....	4
1.2.2 Les publications scientifiques et services associés	5
1.2.3 Informations et ressources pour le grand public, les patients et les soignants.....	6
1.2.4 Les Informations et ressources pour les professionnels de santé.....	7
2. La qualité de l'information Covid19.....	9
2.1 Des données incomplètes, des problèmes de fiabilité.....	9
2.2 L'intégrité scientifique interpellée	11
3. Leçons Covid19 et science ouverte.....	12
3.1 Une dynamique mondiale	12
3.2 Evolution de la communication scientifique en virologie	12
3.3 Questions récurrentes pour la science ouverte.....	14
3.4 Gouvernance sanitaire et <i>data science</i>	15
Conclusion : les métiers de l'information en perspective.....	15
Bibliographie.....	16

Introduction

La science ouverte se déploie désormais depuis plus de 20 ans, revendiquant une science accessible à tous, mobilisable par les acteurs sociaux, économiques, facilitatrice d'autres découvertes scientifiques, transparente pour la société civile. Ce mouvement est indissociable de la vision du bien scientifique commun, de l'éthos du chercheur tel que l'avait décrit Robert K. Merton (Merton, 1973). Mais l'ampleur du mouvement trouve aussi désormais sa source dans la société numérique où les promesses de la science des données (fouilles de données, modélisations, prédictions...) accélèrent les initiatives politiques, notamment pour la Covid19 (OCDE, 2020c). Cette vision n'est pas sans controverse concernant les difficiles équilibres économiques à trouver pour les acteurs directement concernés, notamment pour la diversité des éditeurs œuvrant à la diffusion des résultats scientifiques (Chartron, 2020).

Quoiqu'il en soit, on peut faire l'hypothèse que la Covid19 aura été un accélérateur de la science ouverte et aura montré, qu'en situation de crise, l'ensemble des acteurs pouvaient se mobiliser pour rendre accessible les résultats obtenus mondialement pour la compréhension de ce virus. Ce texte s'attachera à vérifier en partie cette hypothèse. Inversement, on peut aussi faire l'hypothèse que la science ouverte aura permis de faciliter et d'accélérer les résultats scientifiques sur la Covid19. Cette hypothèse est plus difficile à prouver et nécessiterait des études de terrain au plus près des équipes de chercheurs. Nous ne prétendons donc pas d'y répondre.

Cette contribution propose dans un premier temps, de rendre compte de l'explosion des connaissances scientifiques qui ont été partagées dans cette situation de crise pour des publics variés.

Puis nous insisterons sur le fait que cette pandémie a ravivé des questions de fond sur la qualité de l'information dont les effets néfastes de la désinformation. L'accélération des échanges scientifiques a également renouvelé l'enjeu des prépublications avec des dilemmes sur les risques et les opportunités de la circulation de tels résultats.

Enfin, quelles leçons tirer de cette pandémie Covid19 pour le déploiement de la science ouverte en virologie et plus largement ? Nous proposerons de mettre en avant quelques points retenus.

1. Explosion des partages de connaissances sur la Covid19

1.1 Impulsions politiques

Le virus SARS-CoV-2 apparaît en décembre 2019 dans la ville de Wuhan. Le 12 janvier, la Chine partage la séquence génétique du nouveau coronavirus (appelé 2019-nCoV, pour « nouveau coronavirus ayant émergé en 2019 »). L'Organisation Mondiale de la Santé (OMS) déclare la pandémie le 11 mars 2020 en notant qu'il s'agit de la première pandémie causée par un coronavirus.

Dès janvier 2020, la mobilisation politique s'accélère. Plusieurs appels internationaux se succèdent pour un partage mondial des connaissances acquises sur ce nouveau virus. En janvier également, sous l'initiative du Wellcome Trust (UK), 117 organisations (financeurs de recherches, centres de recherche en épidémiologie, éditeurs scientifiques...) signent un engagement commun « *Sharing research data and findings relevant to the novel coronavirus outbreak* »¹. Puis une initiative ralliant 12 pays² dont les Etats-Unis suivra en mars 2020, « *Public Health Emergency Covid-19 Initiative* »³, demandant à tous les éditeurs de mettre en accès libre leurs publications sur la Covid19 et les données associées, de les déposer dans le portail *PubMed Central* et autres répertoires publics majeurs du domaine comme celui de l'organisation mondiale de la santé, the *World Health Organization's COVID data base*, afin de répondre à l'urgence de santé publique. L'appel sera suivi de façon unanime. L'organisation mondiale de la santé et l'Unesco⁴ joueront un rôle majeur pour le rassemblement, le partage, la prise de décision collective afin d'enrayer cette pandémie.

A leur tour, les bibliothèques scientifiques mondiales appelleront à un accès sans barrière à toutes les publications des éditeurs scientifiques utiles à la recherche scientifique sur ce virus⁵.

Si la science ouverte a pu dans bien des cas interroger sur la finalité poursuivie, notamment en SHS (Chartron, 2014), dans le contexte de cette urgence sanitaire, l'adhésion fut quasi-immédiate. Le contexte donne le sens, comme très souvent.

¹ <https://wellcome.org/press-release/sharing-research-data-and-findings-relevant-novel-coronavirus-ncov-outbreak>

² Australia, Brazil, Canada, Germany, India, Italy, Japan, New Zealand, Singapore, South Korea, the UK, and USA

³ <https://www.ncbi.nlm.nih.gov/pmc/about/covid-19/>

⁴ <https://en.unesco.org/news/unesco-mobilizes-122-countries-promote-open-science-and-reinforced-cooperation-face-covid-19> ; <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>

⁵ <https://icolc.net/statement/statement-global-covid-19-pandemic-and-its-impact-library-services-and-resources>

1.2 Inflation et diversité des informations

Un constat majeur est celui d'une très grande profusion et diversité des informations, données, connaissances échangées, et pour des publics différents : chercheurs, professionnels de santé, politiques, grand public... Les initiatives sont menées à la fois par les institutions expertes dans le domaine de l'épidémiologie, des financeurs publics et privés de la recherche, des laboratoires, des universités, des collectifs d'étudiants et de chercheurs, des bibliothèques... L'inflation de sites référencés finit par installer un trop-plein d'informations dont il est souvent difficile de hiérarchiser la valeur et d'en comprendre les éventuelles complémentarités. A cela s'ajoute un prisme international impliquant une lecture dans des périmètres géopolitiques différents.

Nous allons tenter d'éclairer une classification de ces informations dans les paragraphes suivants.

1.2.1 Les données virales factuelles

Un partage fondamental concerne les données virales sur le SARS-CoV2, séquençages de son génome qui permettent aux chercheurs, aux laboratoires de confronter leurs résultats, de contribuer à la mise au point des traitements et des vaccins. Ces séquences génomiques sont déposées dans quelques répertoires ciblés au niveau mondial dont l'enjeu est aussi de les coordonner. Citons 3 organismes majeurs réunis au sein de l'*International Nucleotide Sequence* (INSDC)⁶ travaillant à la coordination et à l'intégration des séquençages déposés sur chacun des répertoires : le NCBI américain, l'EMBL-EBI européen et le DDBJ japonais⁷. Une autre initiative importante concerne le service « *Global initiative on sharing all influenza data* » (GISAID), parallèle aux répertoires publics, créé en 2008 par les scientifiques au niveau international et visant à mieux protéger la reconnaissance individuelle des chercheurs dans leurs partages de données, avec l'intégration d'une licence particulière qui notamment interdit le partage des données avec des utilisateurs sans compte GISAID. Cette base semble être la plus utilisée pour suivre au niveau international l'évolution du virus (Alizon, 2021), l'Institut Pasteur notifie ainsi sur son site : « Les deux séquences complètes des virus prélevés sur deux des premiers cas français ont été déposées le 30 janvier sur la plateforme du «Global initiative on sharing all influenza data » (GISAID), initialement développée pour le partage des séquences et le suivi de l'évolution génétique des virus grippaux, essentiels pour la définition de la composition du vaccin contre la grippe. Un onglet spécial "coronavirus" a été créé pour que la communauté scientifique

⁶ <https://www.insdc.org/>

⁷ <https://www.ncbi.nlm.nih.gov/>; <https://www.ebi.ac.uk/>; <https://www.ddbj.nig.ac.jp/index-e.html>

collabore et avance plus vite. »⁸. Curieux paradoxe pour la science ouverte qui mériterait un approfondissement sur la tension entre la vision d'un « bien commun » et la « compétition scientifique », compétition qui s'affiche comme particulièrement féroce dans le domaine de la biologie (Pearson, 2003).

Concernant toutes ces bases de données de séquençage de génomes, un problème majeur demeure l'intégration et l'harmonisation des données qu'il est intéressant de croiser (Bernasconi & all.,2021). Une ontologie collaborative sur la Covid-19 a été initiée, *The Coronavirus Infectious Disease Ontology* (CIDO)⁹.

1.2.2 Les publications scientifiques et services associés

Comme l'ont défendu plusieurs mobilisations politiques et scientifiques citées au début de l'article, ce sont aussi les publications scientifiques relatives aux résultats de recherche qui ont été mises en libre accès, accessibles par tous les chercheurs, laboratoires de recherche, organismes divers de la santé publique. Des outils d'intelligence artificielle sont conjointement partagés et nécessitent la mise à disposition de données massives et exhaustives pour l'élaboration de modélisations et de prédictions. La base de l'OMS¹⁰ contient ainsi près de 300 000 articles dont 25 000 prépublications, elle est mise à jour quotidiennement. La base PubMed Central, quant à elle, contient 192 445 articles sur la Covid19 dont 96% en libre accès. Les périmètres diffèrent en partie car l'OMS a intégré des ressources de différentes disciplines. PubMed Central mentionne que plus de 50 éditeurs ont répondu à l'appel d'ouverture¹¹.

Des services ont été adossés à ces corpus d'articles régulièrement actualisés : les services de veille qui permettent de diffuser les nouveaux résultats en flux continu, les services de revues systématiques et de méta-analyses qui collectent et produisent des synthèses de connaissances sur une question donnée avec une grande objectivité, définissant les critères d'inclusion et d'exclusion des articles et mettant en évidence les biais potentiels. Un acteur historique du savoir-faire concernant les revues systématiques est le réseau Cochrane¹², organisation indépendante à but non lucratif, spécialisé sur l'information de qualité pour les décisions en matière de santé. D'autres acteurs ont également

⁸<https://www.pasteur.fr/fr/espace-presse/documents-presse/institut-pasteur-sequence-genome-complet-du-coronavirus-sars-cov-2>

⁹ <https://bioportal.bioontology.org/ontologies/CIDO/?p=summary>

¹⁰ <https://search.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/>

¹¹ <https://www.ncbi.nlm.nih.gov/pmc/about/covid-19/#collab>

¹² <https://www.cochrane.org/fr/about-us>

proposé plus récemment ce type de service, comme l'agence ANRS/Maladies infectieuses émergentes¹³, et la documentation centrale des Hospices Civils de Lyon avec le lancement de LittéCovid¹⁴ et MetaEvidence¹⁵. Ce type d'analyse a été mobilisée par plusieurs chercheurs dans le cas très controversé de l'efficacité du traitement par hydrochloréquine¹⁶.

Un dernier type de service concerne la fouille de données par des algorithmes (ou *data-mining*) visant, par exemple, à identifier automatiquement des clusters de travaux ou à répondre à une question particulière en relevant les corrélations dans les textes. Ces services nécessitent des mises en format adaptées et des licences permettant ces usages¹⁷. Il serait intéressant d'en apprécier le développement réel au plus près des équipes en épidémiologie.

1.2.3 Informations et ressources pour le grand public, les patients et les soignants

La Covid19 est certainement la première pandémie qui se conjugue à très grande échelle avec les technologies de réseaux sociaux. Si l'échange collaboratif était une belle promesse, l'inflation de fausses informations circulantes est hélas devenue une réalité inquiétante. L'ampleur est telle que l'OMS, les Nations Unies et d'autres partenaires se sont mobilisés dans une déclaration conjointe «Gestion de l'infodémie sur la COVID-19 : Promouvoir des comportements sains et atténuer les effets néfastes de la diffusion d'informations fausses et trompeuses »¹⁸.

Une priorité est donc de combattre la désinformation sur la Covid19, de lutter contre l'obscurantisme, et remettre les faits et méthodes scientifiques au cœur de l'information diffusée à la société. Difficile de lutter contre de fausses informations qui deviennent très vite virales sur les réseaux sociaux. Le phénomène n'est pas propre aux crises sanitaires mais il devient particulièrement problématique quand il s'agit de s'opposer à la vaccination en mobilisant des rumeurs relayées par activistes et idéologues variés. Faire la différence entre « des faits » et « des opinions » devient un défi pour les sociétés numériques, et cette crise sanitaire aura renouvelé l'enjeu de promouvoir auprès du plus grand nombre

¹³ <https://reacting.inserm.fr/covid-19/covid-19-resources-2/>

¹⁴ <https://rbh.docchu-lyon.fr/littecovid>

¹⁵ <http://www.metaevidence.org/COVID19.aspx>

¹⁶ Effect of hydroxychloroquine with or without azithromycin on the mortality of coronavirus disease 2019 (COVID-19) patients: a systematic review and meta-analysis, <https://www.clinicalmicrobiologyandinfection.com/action/showPdf?pii=S1198-743X%2820%2930505-X>

¹⁷ <https://www.ncbi.nlm.nih.gov/pmc/tools/textmining/>

¹⁸ <https://www.who.int/fr/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation>

une culture scientifique, notamment auprès des jeunes¹⁹. Une stratégie possible est de déployer une communication empruntant les mêmes canaux de diffusion que ceux des influenceurs ; les vidéos YouTube « Canal Detox » de l'INSERM veulent ainsi couper court aux fausses informations et remettre au premier plan les faits scientifiques²⁰.

Renouveler la communication sciences-société est un défi relevé par plusieurs organismes scientifiques qui ont ouvert des sites web dédiés à la Covid19, livrant des explications simples sur la pandémie, ses origines, le contexte historique des précédentes pandémies, l'état des connaissances sur ce virus, l'enjeu des vaccins... L'INSERM²¹, l'institut Pasteur, le CNRS ont poursuivi cet objectif misant sur leur légitimité. De la même façon, les agences de santé publique de chaque pays ont joué un rôle important, comme Santé Publique France²² et ses homologues dans chaque pays. Au plus haut niveau politique, les sites des gouvernements ont également suivi cette priorité pour livrer une information fiable, actualisée à tous les citoyens²³.

Les datavisualisations des données sanitaires relevées au niveau régional, national, européen et mondial ont également largement participé à la transmission des connaissances sur cette pandémie, objectivant les discours, neutralisant certaines rumeurs même si, comme nous le verrons par la suite, les biais de certaines données n'ont pas toujours été précisés lors de leur diffusion, mais décryptés dans un second temps.

1.2.4 Les Informations et ressources pour les professionnels de santé

De nombreux sites diffusent des informations sur les protocoles de diagnostics et de soins pour les professionnels de santé, l'OMS met à disposition des cours en ligne, traduits dans plusieurs langues sur la plateforme OpenWHO²⁴. Des informations plus pointues sur les traitements s'adressent aux médecins comme le service Meta-Evidence-Covid19 qui maintient à jour une méta-analyse des articles relatifs à différents traitements testés sur des patients, l'évolution des décès, les hospitalisations et autres critères que le médecin peut faire varier selon ses préoccupations (cf. figure 1). On voit ainsi

¹⁹ <https://www.clemi.fr/fr/guide-famille/activites-en-famille/cinq-gestes-barrieres-contre-linfodemie.html>

²⁰ https://www.youtube.com/playlist?list=PL525ZU55fXEzw1TTUMi_xkhtq6YRzYvA1

²¹ <https://www.inserm.fr/dossier/coronavirus-sars-cov-et-mers-cov/>

²² <https://www.santepubliquefrance.fr/dossiers/coronavirus-covid-19>

²³ <https://www.gouvernement.fr/info-coronavirus/comprendre-la-covid-19>

<https://www.gouvernement.fr/info-coronavirus/carte-et-donnees>

²⁴ <https://openwho.org/pages/about>

que les 14 articles retenus selon les critères de cette méta-analyse convergent sur le fait que le traitement par chloroquine ne diminue pas les décès et n'induit guère d'améliorations cliniques.

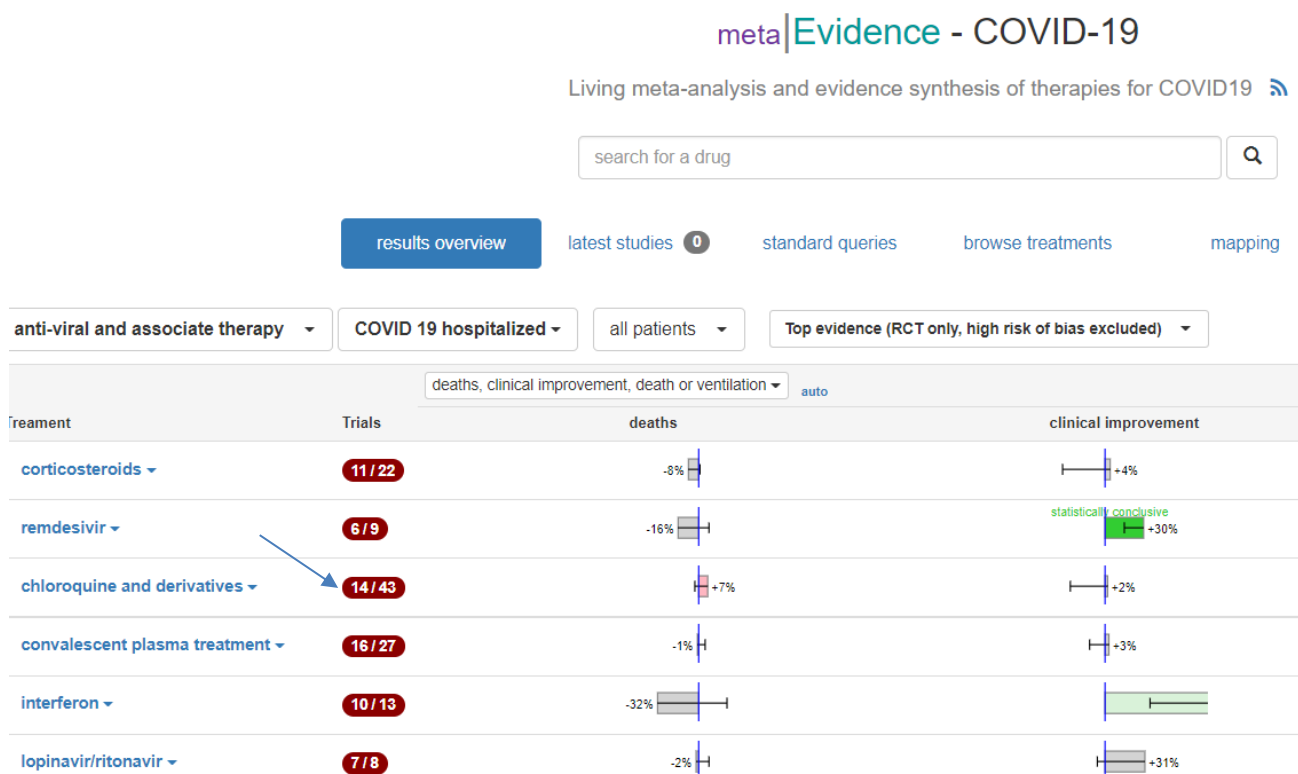


Figure 1 : service Meta Evidence, <http://www.metaevidence.org/COVID19.aspx>
Université de Lyon (UdL/Idex), Hospices Civils de Lyon, CNRS UMR5558 LBBE University Lyon-1.

Par ces différents exemples, nous avons approché la richesse des informations scientifiques livrées en accès ouvert. Le lecteur intéressé à approfondir cette exploration pourra se référer aux deux ressources suivantes : l'inventaire d'Alain Paic à l' OCDE (OCDE, 2020b) et l'inventaire de l'organisation Cochrane²⁵.

Hyper-offre de ressources et de services, implication des experts et émergence de nouvelles initiatives dans chaque pays caractérisent le paysage informationnel Covid19. L'appréciation des usages resterait à faire, notamment pour vérifier l'audience et la reconnaissance respective de tous ces développements. Assiste-t-on à de nouvelles légitimités ou/et les services experts s'en trouvent-ils plus renforcés ?

²⁵ <https://www.cochrane.org/fr/our-evidence/coronavirus-covid-19-resources>

2. La qualité de l'information Covid19

Si l'environnement Covid19 a été marqué par une inflation de données, il n'a pas été exempt de problèmes de qualité de ces données et ceci à plusieurs niveaux. Le problème majeur fut (et demeure) celui de la désinformation, de l'infodémie grandissante. Nous avons déjà évoqué ce point précédemment et souligné l'importance de rétablir des médiations science-société validées, les organismes de recherche en santé publique ont réaffiché cette priorité face à l'ampleur du phénomène.

Les problèmes de qualité ont aussi touché certaines datavisualisations des données épidémiques, très populaires par leur pouvoir performatif mais aussi parfois trop vite élaborées, sans contrôle approfondi de la qualité des données qui les alimentent, c'est un problème classique de la datavisualisation. Données incomplètes, codées différemment, peu comparables ...

2.1 Des données incomplètes, des problèmes de fiabilité

On peut faire l'hypothèse que les données ont rencontré au moins 3 types de problèmes : une dimension organisationnelle sur la collecte des données selon les pays, une dimension politique sur la rétention de certains Etats à transmettre des données fiables, et une dimension technique relative à leur codage, leur standardisation, leur disponibilité dans un système d'information spécifique.

La France dispose de systèmes d'informations de santé développés mais multiples²⁶ dont l'agrégation des données n'est pas toujours facile car les données ne sont pas collectées de la même façon, certains périmètres peuvent être mal couverts, des doublons peuvent se profiler. Une première source concerne les données recueillies puis conservées dans le système d'information SI-DEP, remontées par plus de 600 laboratoires d'analyse publics et privés, quel que soit leur système informatique. Une seconde source concerne les données recueillies puis conservées dans Contact-COVID enregistrées par les médecins généralistes, les enquêteurs sanitaires de la CNAM, Santé publique France et des Agences régionales de santé. Quant aux données hospitalières relatives à l'épidémie de COVID-19, elles sont collectées dans le système d'information SIVIC. Sont également à répertorier les données de SOS Médecins, les données de l'INSEE...

²⁶ <https://www.data.gouv.fr/fr/pages/donnees-coronavirus/>

Le pilotage par la donnée ainsi que la communication publique par les données ne peuvent se déployer sans précaution d'énonciation sur les biais des sources et de leur agrégation. Plusieurs épisodes médiatisés ont marqué ces problèmes qualitatifs lors desquels la communication publique de chiffres a été approximative : le nombre de décès en EHPAD avec des erreurs de saisie, des données incomplètes sur les décès de Santé publique France qui ne couvrent que 3000 communes, soit 77% de la mortalité nationale (cf. explications de l'Insee)²⁷. La crise aura certainement permis de corriger et d'améliorer progressivement les systèmes d'information de santé en France.

A l'international, de nombreux pays ont présenté des données plus problématiques, faute de système d'information en place, faute de volonté ou de coordination politique affirmée. Ainsi, en Inde, le nombre de décès recensés liés à la Covid19²⁸ a été très peu fiable. Selon l'enquête publiée par le Lancet en 2020 (Chatterjee, 2020), parmi les morts enregistrés à l'état civil, seuls 22% ont une cause médicale identifiée et 65% des décès Covid19 proviennent uniquement de 4 Etats (sur 29 en Inde), seules les données des laboratoires de test et des hôpitaux ont été considérées. Les mêmes problèmes de données lacunaires concernant les registres nationaux civils se retrouvent dans d'autres pays comme le Brésil, le Pakistan, le Mexique... (Lloyd-Sherlock & all, 2020). En Europe, le décompte des décès en dehors des hôpitaux a été une source de sous-estimation importante comme en Italie notamment.

Une autre dimension concerne la rétention politique des données Covid19 dans certains pays comme pointé en Russie ou en Chine, la production et la diffusion des données y sont des monopoles d'Etat. Ces pays ont, de façon curieuse, affiché des taux de mortalité très bas à plusieurs périodes de la crise.

Enfin, d'un point de vue technique et informatique, les données épidémiologiques remontées par chaque pays et centralisées dans des répertoires partagés ont présenté également des problèmes : erreurs de dates, codifications différentes, valeurs aberrantes.... Trois répertoires majeurs ont été finement analysés par des chercheurs (Ashofteh, Bravo, 2020) : celui du *World Health Organization*, celui de l'*European Centre for Disease Prevention and Control* et celui du *Chinese Center for Disease Control and Prevention*. La question est centrale pour comparer, suivre et anticiper l'évolution de la pandémie, notamment avec des modèles statistiques. Ces chercheurs ont insisté sur l'enjeu de partager les mêmes standards et de disposer des métadonnées permettant de comprendre le contexte de collecte

²⁷ Insee : Communication sur la mortalité dans le cadre de la pandémie de Covid-19 Mise à disposition de données d'état civil par l'Insee, <https://www.insee.fr/fr/statistiques/4637552?sommaire=4487854>

²⁸ Estimation de 65 288 décès Covid19 en septembre 2020 (Patralekha, 2020)

des données. L'importance de cette documentarisation des données est perceptible sur la plateforme Santé publique où toutes les données sont désormais bien explicitées²⁹.

2.2 L'intégrité scientifique interpellée

La Covid19 a généré une inflation de publications au niveau international, provoquant une difficulté croissante d'évaluations dans les revues scientifiques. Dans un article récent, Pierre Corvol souligne l'ampleur du phénomène (Corvol, 2021):

« Chaque semaine, depuis mai 2020, près de 2 300 articles sont publiés sur la Covid-19. Comme pour toute publication scientifique, les manuscrits sont évalués de façon critique par les pairs (peer-review) et par les éditeurs qui décident de l'acceptation du manuscrit. La participation chinoise a été importante, surtout au début de la crise. Les journaux prestigieux ont pu recevoir jusqu'à une centaine de manuscrits par jour pour en publier moins de 2 ou 3 %. Plusieurs d'entre eux ont demandé aux examinateurs une revue rapide des manuscrits pour permettre une publication express. »

Evaluation trop rapide, manque de filtrage par un comité d'Ethique (comme le recommande *l'International Committee of Medical Journal Editors*)? Quoiqu'il en soit, les retraits d'articles scientifiques se sont multipliés pour cause d'intégrité scientifique, à savoir des résultats scientifiques entachés de fraude aux données, de manipulation des données, de plagiat, de conflit d'intérêt... En un an, le site *retraction watch*³⁰ a recensé 137 retraits d'articles concernant la Covid19, publiés dans des revues scientifiques. Certains retraits ont été particulièrement médiatisés en France, comme les deux articles concernant l'hydroxychloroquine sur le traitement des patients : l'article de Mehra du *Lancet*³¹ et celui du *New England Journal of Medicine*³². La non-efficacité de l'hydroxychloroquine s'appuyait sur des données de 671 hôpitaux associés à la société Surgisphere, données qui posaient de nombreux problèmes d'intégrité, de validité et de transparence.

Le phénomène n'est pas nouveau comme le rappelle J-P. Sculier (Sculier, 2020), mais la course à la reconnaissance scientifique, aux traitements thérapeutiques et les enjeux économiques sous-jacents ont

²⁹ <https://www.data.gouv.fr/fr/pages/donnees-coronavirus/>

³⁰ <https://retractionwatch.com/retracted-coronavirus-covid-19-papers/>

³¹ Mehra MR, Desai SS, Ruschitzka F, Patel AN. RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *The Lancet*. 2020, DOI:[https://doi.org/10.1016/S0140-6736\(20\)31180-6](https://doi.org/10.1016/S0140-6736(20)31180-6)

³² Mehra MR, Desai SS, Kuy S, Henry TD, Patel AN. Retraction: Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. *N Engl J Med*. DOI: 10.1056/NEJMoa2007621.

conduit à une multiplication de mauvaises conduites scientifiques. Plus fondamentalement, c'est la qualité des filtrages de la communication scientifique qui a été questionnée. Au niveau du grand public, même si globalement ces dérives restent malgré tout marginales par rapport à l'ensemble des articles publiés, la légitimité des revues scientifiques a été attaquée, alimentant la défiance ambiante envers la science ainsi que les dérives complotistes de tout genre.

3. Leçons Covid19 et science ouverte

Nous suivons le développement de la science ouverte depuis le début des années 2000, nous aimerions donc insister désormais sur certains points qui ont retenu notre attention dans le contexte de la Covid19.

3.1 Une dynamique mondiale

Le partage des données de séquençage en biologie est un prérequis, impulsé par les politiques publiques internationales et mis en œuvre par les éditeurs lors de la soumission des articles. L'organisation des séquençages peut être différent selon les pays. En France, après un retard accusé, la coopération entre plateformes publiques et laboratoires privés pour le droit à séquençage a progressé, multipliant les capacités de séquençage pour mieux connaître le virus³³. Au final, on constate que le gène de la Covid19 a été séquencé en quelques semaines. Ce sont aussi des milliers de séquences qui sont régulièrement recensées, partagées pour suivre les mutations (processus appelé « la phylogénie »). Toutes les revues concernées ont ouvert leurs publications à la demande des politiques et des scientifiques. Il est ainsi rassurant de voir qu'en cas de pandémie, les coopérations mondiales fonctionnent, que tous les acteurs finissent par se mobiliser. La science ouverte s'est mise en marche, les savoirs ont circulé pour le progrès scientifique, luttant aussi contre l'infodémie galopante.

3.2 Evolution de la communication scientifique en virologie

La pandémie aura, sans nul doute, impacté les pratiques de communication en virologie, concernant les prépublications (*preprints*), versions d'articles scientifiques non encore validées par le comité scientifique d'une revue. Jusque-là assez réfractaires à la circulation de résultats non encore validés par une revue, les chercheurs ont, au contraire, posté massivement leurs articles sur les bases de *preprints* du domaine (medrxiv.org et biorxiv.org principalement). Ainsi, le site Covid19-NIH³⁴ a

³³ <https://lesbiologistesmedicaux.fr/nos-actions/communiqués/872-aventures-françaises-du-séquencage-sars-cov-2>

³⁴ <https://icite.od.nih.gov/covid19/search/>

suivi l'évolution de la proportion de *preprints* dans la littérature scientifique considérée par le NIH (cf. figure 2) et relevé, par exemple, qu'en mai 2020, 32% de ces articles concernaient des prépublications alors que ce taux était de 3% en moyenne sur l'année 2019. Ces prépublications ont également été très citées dans les échanges entre chercheurs (Fraser & all, 2020).

Mais l'association ASAPbio (*Accelerating Science and Publication in biology*), militante pour l'usage des *preprints* en biologie, pointe cependant très clairement le dilemme qui est posé : entre rapidité de diffusion de résultats scientifiques (vs) risque d'une dissémination d'informations fausses pouvant être considérées comme des preuves cliniques établies³⁵.

Par ailleurs, comme évoqué précédemment, les revues n'ont pas toujours été la garantie d'une intégrité scientifique...L'observation échanges a montré que, dans ce domaine, l'évaluation par les pairs pouvait aussi s'engager de façon directe entre chercheurs sur des réseaux tels que Twitter ou ResearchGate (Boudry, Bouchard, 2017).

La Covid19 marquera ainsi de nouvelles articulations, de nouveaux équilibres entre revues et *preprints* dans le domaine de la virologie comme la possibilité de soumettre des *preprints* à bon nombre de revues (qui refusaient jusqu'à présent de considérer des articles ayant circulé au préalable sur Internet), ainsi que des règles resserrées d'évaluation sur la vérification des données comme l'a précisé la revue Lancet après le retrait de l'article sur l'hydroxychloroquine.

³⁵ <https://asapbio.org/preprints-and-covid-19>

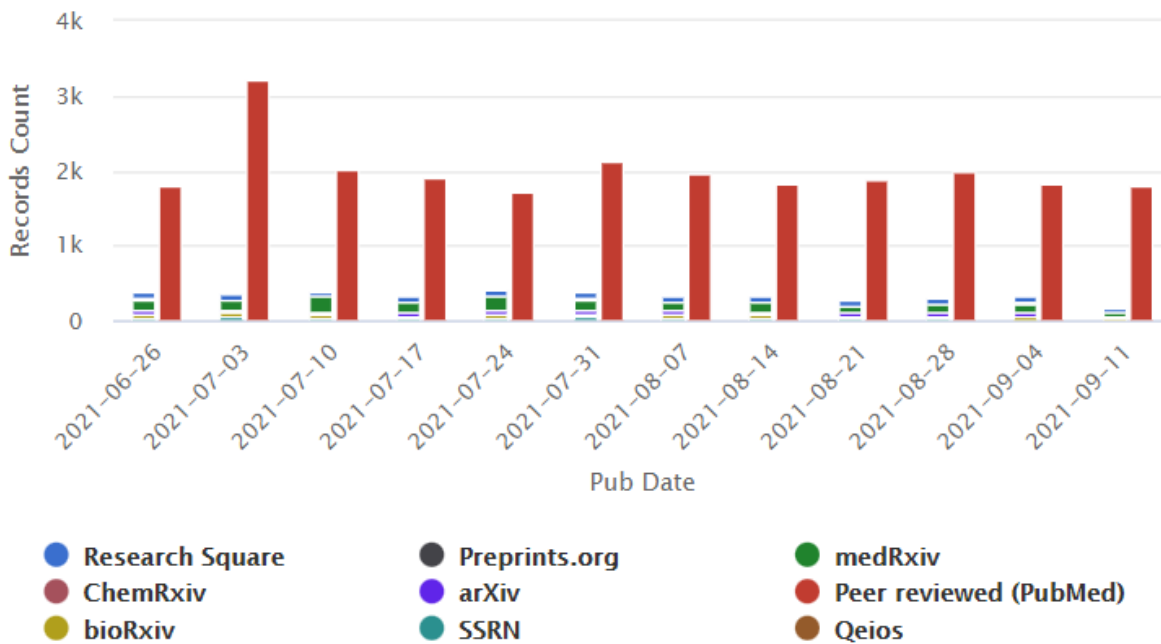


Figure 2 : Nombre de *preprints* par semaine intégrés dans le Portofolio Covid19 du NIH, selon les différentes bases de *preprints*, et comparé à la littérature scientifique de PubMed <https://icite.od.nih.gov/covid19/search/>

3.3 Questions récurrentes pour la science ouverte

Même s'il est très difficile de le démontrer, la libre circulation des données et des résultats scientifiques, a manifestement accéléré le rythme de la recherche pour lutter contre la pandémie.

Mais il n'en reste pas moins que les critiques déjà soulevées sur la vision d'une science qui devrait être ouverte au partage sans limite et sans prise en compte de chaque contexte (Rebouillat, Chartron, 2019), trouvent ici quelques échos. Les enjeux de reconnaissance scientifique dans un contexte très concurrentiel entre équipes pour l'octroi de budgets ainsi que les enjeux économiques sous-jacents aux vaccins sont omniprésents. Nous citerons à nouveau le service GISAID, « *Global initiative on sharing all influenza data* », qui vise à mieux protéger la reconnaissance individuelle des chercheurs, il est apparemment le plus utilisé par les équipes scientifiques. De la même façon, les tensions entre organismes publics et industrie pharmaceutique sur la mise à disposition des données cliniques pour tous les nouveaux médicaments et vaccins³⁶ restent difficiles à lever. Tout ceci renvoie

³⁶ <https://www.who.int/fr/news/item/06-11-2020-who-icmra-joint-statement-on-the-need-for-improved-global-regulatory-alignment-on-covid-19-medicines-and-vaccines>

fondamentalement au modèle économique des acteurs, à leur performance, à leur capacité d'innovation pour la mise au point des traitements et des vaccins contre les virus.

Cette pandémie aura aussi permis de pointer la nécessité de mettre à disposition des informations adaptées à chaque type de public cible. Ainsi, ce n'est pas la mise à disposition « brute » des articles scientifiques des chercheurs qui permet de redonner confiance en la science pour le citoyen mais ce sont bien des reformulations adaptées, synthétiques, argumentées, compréhensibles par tous qui redonnent force au transfert science-société (Chartron, 2018).

3.4 Gouvernance sanitaire et *data science*

La *data science* a été omniprésente dans cette pandémie, acculturant considérablement politiques et société civile à la visualisation des données ; certains y verront même un excès de communication de crise par les données. Mais quand il s'agit d'épidémiologie et de pandémie mondiale, difficile de ne pas mettre les données sanitaires au premier plan pour expliquer, éclairer les décisions, dimension plus performative que toute autre forme de discours. Les modèles prédictifs élaborés à partir des données ont été aussi largement sollicités et ne se sont guère trompés, nous semble-t-il, sur l'anticipation des pics épidémiques.

Toutefois, plusieurs biais sur la qualité des données ont déjà été soulignés et une leçon majeure de la gouvernance sanitaire assistée par la *data science* sera certainement d'avoir pointé les enjeux de disposer de données plus comparables et donc de déployer des politiques affirmées pour la gouvernance des données au niveau international. C'est d'ailleurs l'une des conclusions majeures formulées dans un document de l'OCDE sur un bilan intermédiaire de la crise (OCDE, 2020a).

Conclusion : les métiers de l'information en perspective

Pour conclure, la Chaire d'ingénierie documentaire que j'anime au CNAM, et qui forme des professionnels de l'information, retiendra une lecture particulière de cette pandémie.

Tout d'abord, l'enjeu de l'échange du savoir qualifié pour accélérer le progrès scientifique au niveau mondial a été réaffirmé. Ce savoir repose sur des articles scientifiques, des données contrôlées, des méthodes pertinentes, briques essentielle d'une intégrité scientifique. Les documentalistes scientifiques participent à consolider cette intégrité, contrairement à certaines formes de communication de surface. Durant cette pandémie, les savoir-faire fondamentaux de ces métiers ont

été mis en avant : qualité de la collecte, filtrage, nomenclatures, métadonnées pour donner sens aux données. La science des données ne pourra pas tenir ses promesses sans ces préalables.

Ce sont aussi les traitements intellectuels qualitatifs qui ont été sollicités : les services de revues systématiques et de méta-analyses qui collectent et produisent des synthèses de connaissances existantes sur une question donnée avec le souci d'une neutralisation des biais potentiels, bien différentes des opinions, des rumeurs, des avis trop facilement affirmés.

Le continuum de la donnée, qu'elle soit textuelle ou factuelle, a été souligné et l'enjeu désormais pour les professionnels de l'information s'affirme concernant la maîtrise des techniques d'agrégation, de visualisation, de restitutions sous forme de tableaux de bord (Arruabarrena, Kembellec, Chartron, 2019).

Et pour terminer, malgré la présence d'un organisme international de coordination comme l'OMS, la pandémie a mis en lumière tous les enjeux associés à la qualité, à l'interopérabilité, au partage de standards, à la documentarisation des données pour qu'elles puissent parler...

Dans un contexte d'une data *science* grandissante, la « gouvernance des données » est bien désormais une priorité, un préalable fondamental à la richesse et à l'intelligibilité de toutes ces données.

Bibliographie

Arruabarrena B., Kembellec G., Chartron G. (2019), Data littéraire & SHS : développer des compétences pour l'analyse des données. *CODATA*, Mar 2019, Marne la Vallée, France. [\(hal-03154432\)](#)

Ashofteh, Afshin and Bravo, Jorge M (2020), « A Study on the Quality of Novel Coronavirus (COVID-19) Official Datasets : *Statistical Journal of the IAOS*, vol. 36, no. 2, pp. 243-251, 2020, DOI 10.3233/SJI-200674

Alizon S. (2021), « Variants du SARS-CoV-2 : « Concernant l'utilisation du séquençage en santé publique, la France a du retard » », *The Conversation*, <https://theconversation.com/variants-du-sars-cov-2-concernant-lutilisation-du-sequencage-en-sante-publique-la-france-a-du-retard-154501>

Bernasconi A., Canakoglu A., Masseroli M., Pinoli P., Ceri S. (2021), « A review on viral data sources and search systems for perspective mitigation of COVID-19 », *Briefings in Bioinformatics*, Mar 22;22(2):664-675. doi: 10.1093/bib/bbaa359. PMID: 33348368; PMCID: PMC7799334

Boudry C., Bouchard A. (2017), Réseaux sociaux académiques et diffusion de la production scientifique des chercheurs en biologie/médecine - L'exemple de ResearchGate, *Medecine Sciences (Paris)* 33 (6-7) 647-652 , DOI: 10.1051/medsci/20173306023

Chartron G. (2014), Open access et SHS: controverses. *Revue européenne des sciences sociales*, 52 (1), [10.4000/ress.2658](https://doi.org/10.4000/ress.2658). [sic_00965272](https://doi.org/10.4000/ress.2658)

Chartron G. (2020), Structure et défis économiques de l'édition. Étienne Anheim et Livia Foraison. *L'édition en sciences humaines et sociales. Enjeux et défis*, Éditions de l'EHESS, pp.207-215, [10.4000/books.editionsehess.28973](https://doi.org/10.4000/books.editionsehess.28973). [hal-03154431](https://doi.org/10.4000/books.editionsehess.28973)

Chartron G. (2018). L'Open science au prisme de la Commission européenne. *Education et Sociétés : Revue internationale de sociologie de l'éducation*, De Boeck Supérieur, [10.3917/es.041.0177](https://doi.org/10.3917/es.041.0177). [hal-01888841](https://doi.org/10.3917/es.041.0177)

Chatterjee Patralekha (2020), Is India missing COVID-19 deaths ?, *The Lancet*, Vol. 396, 5 sept.

Corvol P. (2021) , L'envolée des publications scientifiques en temps de Covid-19 Séparer le bon grain de l'ivraie, *Médecine/Sciences 2021 ; 37 : 315-6*

Fraser N., Brierley L., Dey G. , Polka J. K , Pálffy M., Nanni F. , Coates J.A. (2020), Preprinting the COVID-19 pandemic, bioRxiv 2020.05.22.111294; Now published in *PLOS Biology* doi: 10.1371/journal.pbio.3000959

Lloyd-Sherlock P., Sempe L., McKee M. , Guntupalli A. (2020), Problems of Data Availability and Quality for COVID-19 and Older People in Low- and Middle-Income Countries, *The Gerontologist*, Volume 61, Issue 2, Pages 141–144, <https://doi.org/10.1093/geront/gnaa153>

Merton R.K. (1973), *The Sociology of Science: Theoretical and Empirical Investigations*, University of Chicago Press.

OCDE (2020a), « Why open science is critical to combatting COVID-19 », *OCDE Publishing*, Paris, <https://www.oecd.org/coronavirus/policy-responses/why-open-science-is-critical-to-combatting-covid-19-cd6ab2f9/>

OCDE (2020b), coordonné par Paic A., Open science initiatives related to the COVID-19 pandemic, répertoire actualisé, <https://community.oecd.org/docs/DOC-172520>

OCDE (2020c), Using artificial intelligence to help combat COVID-19.

Pearson, H. (2003), Competition in biology:It's a scoop!. *Nature*, <https://doi.org/10.1038/news031124-9>

Rebouillat V., Chartron G. (2019), Services de gestion et de partage des données de recherche : ce qu'en pensent les chercheurs. *12ème Colloque international d'ISKO-France : Données et mégadonnées ouvertes en SHS : de nouveaux enjeux pour l'état et l'organisation des connaissances?*, Oct 2019, Montpellier, France. [hal-02307085](#)

Sculier J.-P. (2020), Fraudes et plagiats : une situation pas si rare, *Revue Médicale de Bruxelles*, Vol.41-6, décembre 2020