COVID-19 et (dés)information : l'*open data* face à ses limitations

Robert Viseur[1]

1. Service TIC, FWEG, UMONS
17 place Warocqué, B-7000 Mons, Belgique
robert.viseur@umons.ac.be

RÉSUMÉ. *La COVID-19 a mis en évidence l'importance du numérique dans la lutte contre une pandémie. En particulier, la France s'est distinguée par la diversité des jeux de données publiés en open data. Cependant, elle n'a pas été épargnée par les polémiques relatives à leur qualité et à leur interprétation. Sur base d'un ensemble d'études de cas réutilisant les données du SPF (Santé publique France) et de la DREES (Direction de la Recherche, des Études, de l'Évaluation et des Statistiques), nous analysons la qualité des données, en nous basant sur le concept de pertinence des représentations de Reix et al. (2011), puis les obstacles à leur réutilisation. Nous identifions des problèmes classiques en analyse de données (p. ex. difficultés d'interprétation liées au volume ou à la sémantique des données). Nous analysons ensuite les conséquences sur la réutilisation des open data de deux dimensions de ces systèmes d'information : (1) le développement itératif des systèmes d'information de santé utilisés dans le cadre de la pandémie et (2) le fonctionnement en silos de ces systèmes d'information. Dès lors, nous identifions comment les défauts de pertinence peuvent entraîner un défiance ou un mésusage de ces données.*

ABSTRACT. *COVID-19 highlighted the importance of digital technology in the fight against a pandemic. In particular, France stood out for the diversity of the datasets published in open data. However, it has not been spared from controversies regarding their quality and interpretation. Based on a set of case studies reusing SPF and DREES data, we analyse the quality of the data, based on the concept of relevance of representations by Reix et al. (2011), and then the obstacles to their reuse. We identify classic problems in data analysis (e.g., interpretation difficulties related to data volume or semantics). Then we analyze the consequences for open data reuse of two dimensions of these information systems: (1) the iterative development of health information systems used in the pandemic and (2) the siloed operation of these information systems. Therefore we identify how relevance flaws can lead to distrust or misuse of these data.*

Mots-clés : open data, COVID-19, pertinence des représentations.

KEYWORDS: *open data, COVID-19, data quality.*

1. Contexte

En décembre 2019 démarrait en Chine une épidémie de pneumonie attribuée à un nouveau coronavirus baptisé SARS-CoV-2. Parmi les moyens numériques, les *open data*, c'est-à-dire des informations qui ont été rendues techniquement et légalement disponibles à des fins de réutilisation (Lindman & Tammisto, 2011), se sont rapidement dégagées comme un instrument facilitant le suivi de la pandémie. Cependant, si Lobre et Lebraty (2012) reconnaissent que « *la discussion offre une chance de réinterpréter les données de manière efficace et fiable* », ils soulignent aussi les risques liés à la « *manipulation de l'information* ». Un élément nous semble par ailleurs sous-étudié dans le cadre de cette pandémie : les conséquences, sur la qualité de l'information, de la politique de publication des *open data*. Même si la France est apparue comme un bon élève en matière d'*open data* COVID-19 (Viseur, 2021 ; Nikiforova, 2020), des voix se sont aussi élevées pour dénoncer la lenteur de publication des données relatives à la pandémie (Chignard, 2021). La politique en matière d'*open data* a été stimulée par l'activité de collectifs (p. ex. OpenCOVID et « *Data against Covid* ») procédant méthodiquement à l'encodage de nouveaux jeux de données sur base des rapports officiels, publiant des tableaux de bord permettant un suivi de la pandémie et interpellant régulièrement le gouvernement pour qu'il publie davantage de données (Goëta, 2022 ; Chignard, 2021 ; Bothorel et al., 2020). Une fois publiées, ces *open data* peuvent donc faciliter l'information du public. En France, des sites comme CovidTracker[1] sont progressivement devenus incontournables pour le suivi médiatique de la pandémie (Goëta, 2022 ; Bothorel et al., 2020). Malgré ces efforts de transparence, la pandémie de la COVID-19 a été associée à une « *prolifération de théories du complot* » (Bottemanne, 2022). Cette recherche s'attaque donc à la compréhension des facteurs liés à la qualité de ces *open data* expliquant pourquoi, dans un pays ayant joué la carte de la transparence, le suivi de la pandémie n'a pas échappé aux polémiques constatées ailleurs.

Les organisations prennent leurs décisions sur la base de représentations de la réalité (p. ex. tableau de bord de gestion). Dès lors, en tant que représentation d'une réalité, ces *open data* COVID-19 sont-elles « *pertinentes* » ? La « *pertinence des représentations* », analysée par Reix et al. (2011), renvoie à leur utilisation. Est pertinent « *ce qui convient, ce qui est approprié à l'action* » (Reix et al., 2011). Huit critères de qualité permettent, selon Reix et al. (2011, de juger la pertinence des représentations : (1) l'accessibilité (modalités de recherche et d'accès) ; (2) l'exactitude, qui concerne l'absence de bruit (les représentations devraient idéalement échapper au risque de deuxième espèce relatif à la prise en compte d'évènements provoqués par des variations aléatoires dues à des imperfections de la fonction d'information) ; (3) l'exhaustivité, qui fait référence à la complétude de la représentation (tous les changements d'état significatifs du réel sont couverts par les représentations ; cf. risque de première espèce) ; (4) la fiabilité (confiance en la source ; dépendance aux risques de 1ère et de 2ème espèce) ; (5) le degré de finesse, qui concerne la précision de la représentation (niveau de détail ou intervalle de variation) ; (6) l'actualité (fraîcheur), (7) la ponctualité (respect des échéances) ; et (8) la forme (données, dessins, images fixes ou animées…). Cette question de la pertinence des représentations s'inscrit dans un questionnement plus général sur le poids occupé par les chiffres dans notre société. « *On peut débattre de tout sauf des*

[1] Cf. https://covidtracker.fr/.

chiffres » affirmait ainsi le gouvernement français lors d'une campagne en faveur de la vaccination, usant ainsi « *des chiffres comme argument d'autorité* » (Goëta, 2022). Dans une série de cours au Collège de France donnés entre 2012 et 2014, Alain Supiot (2015) critique notamment le mésusage des indicateurs, devant être interprétés, « *substituant la carte au territoire et la réaction à l'action* », congédiant « *le vocabulaire de la démocratie politique au profit de celui de la gestion* », ayant pour effet de se couper de la complexité du réel.

2. Données et analyse des résultats

Les *open data* fournies par l'état français le sont par Santé Publique France (SPF) via le portail data.gouv.fr. Ces données incluent les « *Données hospitalières relatives à l'épidémie de COVID-19* » (cf. [url]), les « *Données relatives aux résultats des tests virologiques COVID-19* » (cf. [url]) et les « *Données relatives aux personnes vaccinées contre la Covid-19* » (cf. [url]). La DREES (Direction de la Recherche, des Études, de l'Évaluation et des Statistiques) a par la suite publié des fichiers constitués des appariements entre ces trois sources de données (cf. [url]). Compte tenu de leur mention fréquente sur les réseaux sociaux francophones, les *open data* américaines relatives aux effets secondaires présumés des vaccins, publiées au sein du VAERS (*Vaccine Adverse Reporting System*), ont également été exploitées. Plusieurs expérimentations sur base de ces données ont été réalisées entre août 2021 et février 2022. Elles suivaient en pratique les polémiques médiatiques, notamment relatives à la comptabilisation du nombre total de décès, aux causes réelles des décès, à l'efficacité des vaccins, à leur innocuité et à l'ampleur des variations du nombre d'hospitalisations (notamment chez les enfants). À l'aide des critères de « *pertinence des représentations* » de Reix et al. (2011), nous avons procédé à l'analyse des données mobilisées dans les cas de réutilisation. Le Tableau 1 distingue, pour chacun des huit critères, les données relatives aux tests, aux hospitalisations, aux décès et aux vaccinations.

Tableau 1. Données COVID et pertinence des représentations.

	(a) Tests	**(b) Hospitalisations**	**(c) Morts**	**(d) Vaccinations**
(1) Accessibilité	Pas d'accès (01/2022) au pourcentage de tests positifs à Omicron.	Pas d'*open data* sur le statut vaccinal (S1 2021) ou sur le statut de comorbidité.	Plus d'accès aux statistiques de décès dans les EHPAD (cf. [url]).	Pas d'accès en France, au contraire des USA (VAERS), aux données brutes relatives aux signalements d'effets secondaires présumés.
(2) Exactitude	Positivité des tests influencée par la capacité ou par la politique de tests.	Pas de distinction entre les hospitalisations « pour COVID » et « avec COVID ».	Pas de distinction entre les morts « pour COVID » et « avec COVID ».	Positivité (tests) des populations vaccinés influencées par les politiques de tests différenciées entre vaccinés et non vaccinés (principe du pass sanitaire).
(3) Fiabilité	Problèmes liés à la sémantique des fichiers *open data*.	Problèmes liés à la sémantique des fichiers *open data* (total des hospitalisations, hospitalisations conventionnelles, soins critiques, réanimations...).	Difficulté à comptabiliser, et distinguer, les morts en hôpital, en EHPAD et à domicile.	Régularité des remontées d'effets secondaires présumés liés aux vaccins (complexité du formulaire de signalement de l'ANSM, actes militants liées aux mouvements antivax...).
(4) Exhaustivité	Influence du choix de technologies de tests comptabilisés (auto-tests, salivaires, antigéniques, RT-PCR).	Indisponibilité, réparée au fil du temps, de l'âge, du statut vaccinal... et du statut de comorbidité.	Comptabilisation séparée des morts en hôpital ou des morts en EHPAD.	Absence d'indicateur lié à l'immunité naturelle (primoinfections).
(5)	Publication d'indicateurs	Publication d'indicateurs	Publication d'indicateurs	Publication d'indicateurs

	(a) Tests	(b) Hospitalisations	(c) Morts	(d) Vaccinations
Finesse	globaux ou d'indicateurs par critère (p. ex. tranche d'âge et statut vaccinal).	globaux ou d'indicateurs par critère (p. ex. tranche d'âge et statut vaccinal).	globaux ou d'indicateurs par critère (p. ex. tranche d'âge et statut vaccinal).	globaux ou d'indicateurs par critère (p. ex. tranche d'âge).
(6) Actualité	Irrégularités dans les remontées de cas (p. ex. *weekends* et reprise des erreurs).	Problème du décalage temporel entre les cas et les hospitalisations (prévisions) et de la variation des délais au fil du temps (variants).	Problème du décalage temporel entre les cas et les décès (prévisions) et de la variation des délais au fil du temps (variants).	Accès aux informations sur les effets secondaires présumés ou confirmés après traitement (ANSM).
(7) Ponctualité	Retards sur le criblage pour le suivi du variant Omicron.		Rattrapage sur les morts en EHPAD (cf. [url1] et [url2]).	
(8) Forme	Focalisation sur des indicateurs globaux, graphiques et spectaculaires (p. ex. exponentielle du nombre de cas).	Utilisation d'échelles permettant d'exagérer les phénomènes (p. ex. hospitalisations pédiatriques).	Focalisation sur des indicateurs globaux.	Focalisation sur des indicateurs globaux (p. ex. graphique de l'évolution de la couverture vaccinale globale).

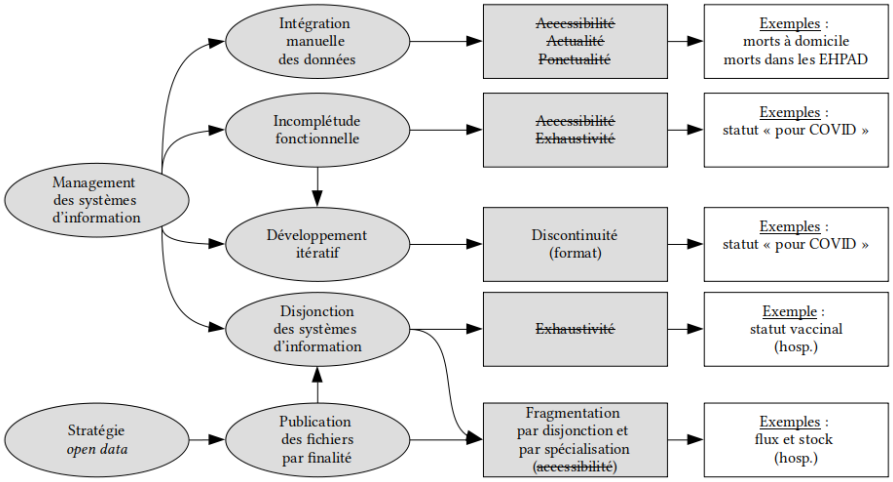

Figure 1. Construction des SI et pertinence des représentations.

Les lenteurs de publication s'expliquent en partie par la construction du système d'information de santé en France. Ronai (2021) distingue ainsi système d'information orienté statistique, avec une filière administrative (INSERM) et une autre sanitaire (INSEE), et système d'information en temps réel. Les *open data* publiées par Santé Publique France (SPF) dépendent essentiellement du second, constitué de trois systèmes d'information : SI-VIC, SI-DEP et VAC-SI (Guillotin et al., 2021 ; Ronai, 2021 ; Bothorel et al., 2020). Créé en 2016, déployé initialement pour répertorier les victimes des attentats de Paris de novembre 2015, SI-VIC sert au départ de système d'information pour le suivi des victimes d'attentats et de situations sanitaires exceptionnelles. Ses fonctionnalités ont été étendues, de manière itérative, pour les besoins liés à la pandémie. Cependant, SI-VIC ne comptabilisait que les décès survenus à l'hôpital et excluait les morts à domicile ou en maison de retraite (soit 44 % des morts de la COVID-19 ; cf. [url]). Ces derniers ont finalement pu être comptabilisés via un quatrième système (Voozanoo) dédié aux EHPAD (Bothorel et al., 2020). SI-DEP, développé par l'AP-HP, disponible depuis juin 2020, contient les résultats des tests PCR et des tests antigéniques. Il est

complété par Contact-COVID qui est dédié au suivi des cas positifs et à la conduite des enquêtes sanitaires permettant la remontée des chaînes de contamination. VAC-SI, développé par le CNAM, disponible à partir du 4 janvier 2021, contient les données relatives à l'identité du patient et à la vaccination.

Si l'existence des systèmes d'information offre globalement les bénéfices d'une intégration rapide des données collectées sur le terrain, ce que la Belgique a par exemple peiné à réaliser (Viseur, 2021), leur développement progressif, itératif, conduit à des défauts d'exhaustivité et à des discontinuités dans les formats de données (cf. Figure 1). De plus, le manque d'intégration entre systèmes d'information conduit à la publication de jeux de données reflétant les silos dont ils sont issus. Les données pour un même fichier sont donc relatives, soit aux hospitalisations et aux décès (dans les hôpitaux), soit aux tests, soit aux statuts vaccinaux. Premièrement, il en résulte des angles morts. Les décès en EHPAD ne sont ainsi pas comptabilisés dans les décès repris dans SI-VIC. Deuxièmement, les croisements entre statut vaccinal et hospitalisation sont difficiles. Au niveau des *open data*, le croisement entre les tests, les hospitalisations et les décès n'a été permis que tardivement (c'est-à-dire début août 2021) suite à la publication par la DREES de fichier d'appariement (cf. [url]) permettant par exemple d'analyser le risque d'hospitalisation ou de décès en fonction du statut vaccinal. Cette fragmentation est renforcée par la stratégie de publication en *open data*, où chaque fichier reflète une finalité particulière. L'analyse d'un phénomène (p. ex. hospitalisations pédiatriques) va donc nécessiter le croisement de plusieurs fichiers, ce qui peut décourager certains médiateurs (*gatekeepers*), en renforcer d'autres, sans que leur professionnalisme ne soit nécessairement au rendez-vous.

À ces difficultés s'ajoutent des problèmes, classiques en analyse de données (Bronner, 2013 ; Baillargeon, 2006), relatives à trois formes de complexité liées (1) au volume de données (p. ex. taille des jeux de données), (2) à la compréhension des données (p. ex. sémantique des données) et (3) à l'interprétation des données (p. ex. signification des indicateurs tels que la létalité, méconnaissance des processus de validation des données brutes et confusion entre corrélation et causalité ; cf. effets secondaires présumés des vaccins par exemple). En particulier, certains problèmes de fiabilité se sont révélés imputables à des problèmes de sémantique des fichiers, handicapant, d'une part, la compréhension du sens associé aux données réutilisées, d'autre part, le croisement des données entre fichiers complémentaires. Ces incohérences ponctuelles dans la sémantique des fichiers portent notamment sur les capacités en réanimation. La Cour des comptes a ainsi publié en juillet 2021 un rapport intitulé « *Les soins critiques* » dans lequel le problème a été décrypté (cf. [url]). Le rapport note la confusion opérée par Santé Publique France (SPF) entre les réanimations et les soins critiques ainsi qu'une incertitude quant, d'une part, aux capacités maximales prises en référence, d'autre part, à la part des lits affectés aux patients COVID (pp. 57-60).

3. Conclusion

Au delà des problèmes classiques en analyse de données, cette recherche a permis d'en dégager de plus spécifiques liés, d'une part, à la temporalité du phénomène, d'autre part, à l'intégration des systèmes d'information de santé. La rapidité et la nouveauté du phénomène ont conduit, d'une part, à un développement dans l'urgence des systèmes d'information, d'autre part, à une mise à mal du rythme

de validation des connaissances. Si l'amélioration continue de la qualité est favorisée par les pratiques collaboratives qui accompagnent souvent l'*open data*, elle peut aussi donner l'impression à un public moins informé d'une piètre fiabilité des données publiées, voire d'une volonté de tromper. La complexité d'interprétation des données se trouve accrue par la nécessité d'identifier, puis de croiser, lorsque cela est possible, les différents jeux de données indispensables à l'analyse.

4. Références

Baillargeon, N.. (2006). Petit cours d'autodéfense intellectuelle. Lux.

Bothorel, É., Combes, S., Vedel, R. (2020). Mission BOTHOREL : pour une politique publique de la donnée. Conseil général de l'économie.

Bottemanne, H. (2022). Théories du complot et COVID-19: comment naissent les croyances complotistes?. L'encephale.

Bronner, G. (2013). La démocratie des crédules. PUF.

Chignard, S. (2021). L'open data de crise : entre mobilisation citoyenne et communication gouvernementale. Enjeux numériques, n°14, juin 2021, pp. 73-77.

Cour des comptes (2021). Les soins critiques. Communication à la commission des affaires sociales du Sénat. Cour des comptes, juillet 2021.

Goëta, S. (2022). La pandémie de Covid-19: un moment charnière pour l'ouverture des données en France. Statistique et Société.

Guillotin, V., Lavarde, C., Savary R.-P. (2021). Crises sanitaires et outils numériques : répondre avec efficacité pour retrouver nos libertés. Rapport d'information au Sénat, n°673, 03 juin 2021.

Lindman, J., & Tammisto, Y. (2011). Open Source and Open Data: Business perspectives from the frontline. In : IFIP International Conference on Open Source Systems (pp. 330-333). Springer, Berlin, Heidelberg.

Lobre, K., & Lebraty, J. F. (2012). L'open Data: nouvelle pratique managériale risquée?. Gestion 2000, 29(4), 103-116.

Nikiforova, A. (2020). Timeliness of Open Data in Open Government Data Portals Through Pandemic-related Data: a long data way from the publisher to the user. In 2020 Fourth International Conference on Multimedia Computing, Networking and Applications (MCNA), pp. 131-138. IEEE.

Reix, R., Fallery, B., Kalika, M., & Rowe, F. (2011). Systèmes d'information et management des organisations. Vuibert.

Ronai, M. (2021). La construction d'un système d'information épidémiologique. Enjeux numériques, n°14, juin 2021, pp. 62-72.

Supiot, A. (2015). La gouvernance par les nombres. Fayard.

Viseur, R. (2021). Open data in digital strategies against COVID-19: the case of Belgium. 17th International Symposium on Open Collaboration. September 2021. Pages 1-10.