

[Atelier 03 – Données de santé]

*« Analyse des limitations à la réutilisation
des open data COVID-19 en France »*

Robert Viseur
Chargé de cours

Journée de recherche RIPCO
*« Comportements de santé, santé des personnes et des organisations :
ce que nous apprend la pandémie COVID-19 »*

Distanciel – 17 juin 2022

Contexte

Débuts de la pandémie de la COVID-19 en décembre 2019 à Wuhan (Dong et al., 2020) puis mesures de confinements dès le 23 janvier 2020 (Lau et al., 2020).

Mesures de confinement similaires en Europe à partir de mars 2020 (Saglietto et al., 2020 ; Desson et al., 2020) dans un contexte de forte pression hospitalière et de (risque de) pénurie d'équipements médicaux (p. ex. visières ; Viseur et Charleux, 2021).

Recours au numérique (profilage, *tracing*, *tracking*, passeports...) en Asie (Viseur, 2021a). Exploitation en Europe des applications de *proximity tracing* (Viseur, 2021a) et de l'*open data* (Nikiforova, 2020).

Réactivité de la France en matière d'*open data* (Nikiforova, 2020), avec de belles réalisations à la clef (p. ex. [CovidTracker](#)), ce qui n'a pas épargné le pays de la polémique (p. ex. hospitalisations pour ou avec COVID) et de la désinformation (p. ex. négation des hospitalisations et exagération des effets secondaires des vaccins).

Question de recherche

Cette recherche s'appuie sur une démarche inductive et itérative d'exploration des jeux de données fournis principalement par les organismes français de santé publique.

Quelles sont les caractéristiques des *open data* publiées en France ?

Quelles informations permettent-elles de dégager ?

Quels problèmes ces *open data* présentent-elles sur le plan de la pertinence des représentations (Reix et al., 2011) ?

Au delà des difficultés classiques liées à l'interprétation des données et des problèmes de désinformations relevés mondialement au cours de la pandémie, quelles sont les caractéristiques des *open data* COVID françaises susceptibles d'alimenter des polémiques ou des désinformations ?

Revue de littérature #1

Concept de « *gouvernance par les nombres* » (Supiot, 2015) :

- Constat du passage d'un gouvernement par la loi (« *subordination* ») à une gouvernance par les nombres (« *programmation* »).
- Usage normatif des indicateurs « *substituant la carte au territoire et la réaction à l'action* ».
- Problématique spécifique des « *indicateurs-objectifs* ».

Concept de « *pertinence des représentations* » (Reix et al., 2011) :

- Prise de décision sur base de représentations de la réalité, obtenues au travers de fonctions d'information, alimentées par la mesure de signaux.
- Question de la pertinence de ces représentations et proposition d'une liste de critères de pertinence : accessibilité, exactitude, fiabilité, exhaustivité, finesse, actualité, ponctualité et forme.

Revue de littérature #2

Problèmes connus d'interprétation de données (Bronner, 2013 ; Baillargeon, 2005) : biais de confirmation, confusion entre corrélation et causalité... y compris dans le contexte des *open data* (Lobre et Lebraty, 2012)... et en particulier dans un contexte favorable à la désinformation voire au complotisme (Bottemanne, 2022).

Données extraites des systèmes d'information COVID en France :

- Disponibilité de 3 SI principaux (Guillot et al., 2021 ; Ronai, 2021) :
 - SI-VIC,
 - SI-DEP,
 - VAC-SI.
- Construction dans l'urgence pour répondre à une contrainte de suivi en temps réel de la pandémie et, pour SI-VIC, au départ d'une application existante.

Données et méthodologie

Focalisation sur les données publiques fournies par les administrations et les agences de santé publique, c'est-à-dire en France par le [SPF](#) (Santé Publique France), via le portail data.gouv.fr (>200 jeux de données COVID au 17/06/2022), et la [DREES](#) (Direction de la Recherche, des Études, de l'Évaluation et des Statistiques).

Complément par les données du [VAERS](#), organisme co-sponsorisé aux États-Unis par le [CDC](#), souvent mentionnées par les milieux hostiles à la vaccination COVID.

Traitement des *open data* du SPF et de la DRESS sous [LibreOffice.org Calc](#) (tableur) et de celles du VAERS avec [Python](#) / [SQLite](#) et LibreOffice.org Calc.

Mise en œuvre de cas de réutilisation entre août 2021 et février 2022 (souvent en fonction des sujets polémiques émergents) : calcul de profil de risque en fonction de la tranche d'âge, calcul de létalité / de mortalité, calcul des durées moyennes d'hospitalisation, production de statistiques d'incidents présumés de vaccination...

Résultats (exploitation)

Disponibilité d'un vaste ensemble de jeux de données, dont la quantité (p. ex. publication régulière de nouveaux fichiers) et la qualité (p. ex. ajout de nouvelles données aux fichiers) évoluent positivement au fil du temps (mise à jour).

Vigueur des pratiques d'*open science* (Fecher et Friesike, 2014) : réutilisation d'*open data* et publication massive de *preprints* (Fraser et al., 2021 ; Vlasschaert et al., 2020).

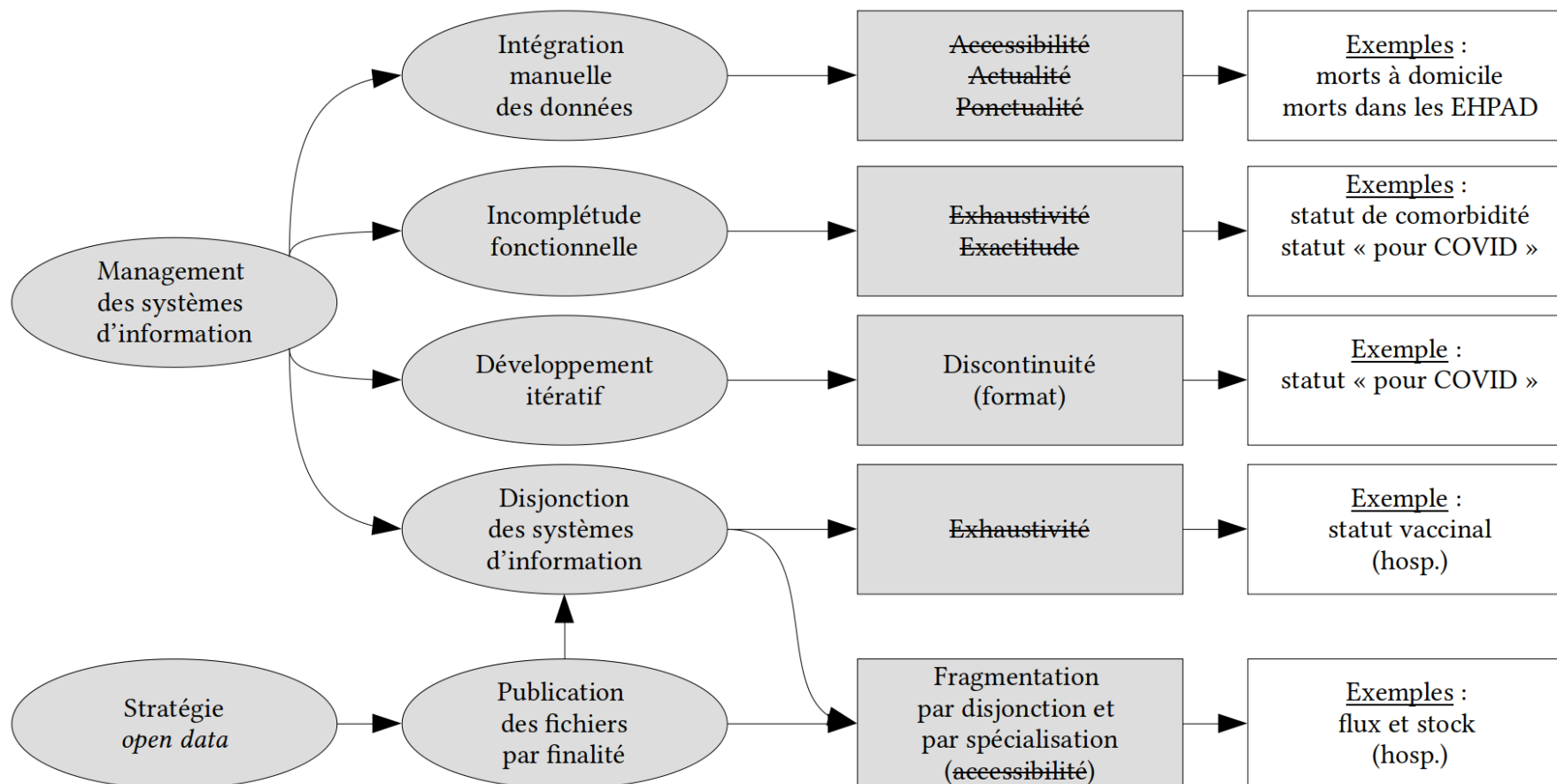
Grande diversité de réutilisateurs (scientifiques, ingénieurs, journalistes...). Émergence de nouveaux *gatekeepers* reconnus pour leur fiabilité (p. ex. [Germain Forestier](#) et [Guillaume Rozier](#)) mais également par des collectifs militants (p. ex. [OpenVAERS](#)).

Obstacle à la réutilisation dû à la fragmentation de l'information. Observation de cas classiques de mauvaise compréhension des données brutes (Lobre et Lebraty, 2012), telles que le biais de confirmation (p. ex. létalité vs mortalité), basées sur des calculs partiels, conduisant parfois à un rejet de la fiabilité des *open data*.

Résultats (pertinence des représentations)

	(a) Tests	(b) Hospitalisations	(c) Morts	(d) Vaccinés
(1) Accessibilité	Pas d'accès (01/2022) au pourcentage de tests positifs à Omicron.	Pas d' <i>open data</i> 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 (volume) relatives aux signalement 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 population 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 <i>open data</i> .	Problèmes liés à la sémantique des fichiers <i>open data</i> (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) Finesse	Publication d'indicateurs globaux ou d'indicateurs par critère (p. ex. tranche d'âge et statut vaccinal).	Publication d'indicateurs globaux ou d'indicateurs par critère (p. ex. tranche d'âge et statut vaccinal).	Publication d'indicateurs globaux ou d'indicateurs par critère (p. ex. tranche d'âge et statut vaccinal).	Publication d'indicateurs globaux ou d'indicateurs par critère (p. ex. tranche d'âge).
(6) Actualité (délai)	Irrégularités dans les remontées de cas (p. ex. <i>weekends</i> 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 (échéance)	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 → indicateur-objectif).

Résultats (systèmes d'information)



Conclusion

Comparativement à la Belgique (Viseur, 2021b), meilleure accessibilité et plus grande exhaustivité des données COVID-19.

Problématiques connues de construction itérative et d'intégration des systèmes d'information (Reix et al., 2011 ; Laudon et Laudon, 2013) mais contrainte nouvelle de temporalité de la pandémie, dès lors exigence de rapidité du processus de développement.

Mise en évidence d'une aggravation de ces difficultés d'exploitation et d'interprétation des données liées au manque d'informatisation, à l'incomplétude fonctionnelle, au mode de développement itératif et au manque d'intégration des SI.

Émergence de nouveaux *gatekeepers* mis en capacité par ces données (p. ex. [Guillaume Rozier](#)) mais aussi de collectifs militants (p. ex. [Côté Science](#), [Covidrationnel](#) et [Reinfocovid](#)), inspirés par les principes de l'*open science*, conduisant à une médiatisation de conflits parfois violents (réinterprétation vs polarisation).

Ce support de présentation est diffusé sous licence CC-BY-ND.

Université de Mons
Faculté Warocqué d'économie
et de gestion - Service TIC
Place Warocqué, 17
B-7000 Mons

Tél. : +32.65.373.201

www.umons.ac.be
info.warocque@umons.ac.be

Plus d'information...

Dr Ir Robert VISEUR
Chargé de cours

Tél. : +32.65.374.054
robert.viseur@umons.ac.be