Open data in digital strategies against COVID-19: the case of Belgium

Robert Viseur
University of Mons
robert.viseur@umons.ac.be

ABSTRACT

COVID-19 has highlighted the importance of digital in the fight against the pandemic (control at the border, automated tracing, creation of databases...). In this research, we analyze the Belgian response in terms of open data. First, we examine the open data publication strategy in Belgium (a federal state with a sometimes complex functioning, especially in health), second, we conduct a case study (anatomy of the pandemic in Belgium) in order to better understand the strengths and weaknesses of the main COVID-19 open data repository. And third, we analyze the obstacles to open data publication. Finally, we discuss the Belgian COVID-19 open data strategy in terms of data availability, data relevance and knowledge management. In particular, we show how difficult it is to optimize the latter in order to make the best use of governmental, private and academic open data in a way that has a positive impact on public health policy.

CCS CONCEPTS

• **Collaborative and social computing and tools**;

KEYWORDS

open data, COVID-19, open science, knowledge management, health

ACM Reference Format:

Robert Viseur. 2021. Open data in digital strategies against COVID-19: the case of Belgium. In *17th International Symposium on Open Collaboration (OpenSym 2021), September 15–17, 2021, Online, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3479986.3479988

1 INTRODUCTION

In December 2019, an outbreak of pneumonia attributed to a new coronavirus named SARS-CoV-2 started in Wuhan City (capital of Hubei Province, China) [9]. This 2019 coronavirus disease (or COVID-19) led to the strict confinement of Wuhan City by January 23, 2020 and most major cities in Hubei by January 24 [17]. Although locally effective, these confinement measures did not prevent the disease from spreading to mainland China and several dozen other countries, leading the WHO to declare the COVID-19 epidemic a pandemic on March 11, 2020. In Europe, the disease hit Italy hard. Despite the gradual implementation of confinement from March 9, the health care system there quickly reached its limits (due in particular to the limited capacity of intensive care units) while other European countries experienced a similar progression of the virus (with a few days or weeks delay) and implemented confinement measures in turn [30]. In Belgium, confinement was imposed as of March 18, 2020, while a second confinement was implemented as of November 2, 2020 after a relaxation of the measures during the summer vacations. Digital technology was part of the arsenal of responses to fight the spread of the virus [34]: profiling of citizens at the borders, identification of contact cases (tracing), cluster analysis, development of big open data... In this difficult health context, some countries published open data related to the COVID-19 as from the first weeks of the pandemic [23]. Open data was also quickly pointed out as a necessary tool to monitor and anticipate the spread of the virus [35].

The unavailability of open data in Belgium was identified at the beginning of the pandemic as a hindrance to the work of scientists[1]. The publication of open data files is the responsibility of Sciensano, a public institution created on February 25, 2018. The unavailability of certain raw data has led to protests from the academic world, sometimes relayed by the press or by elected officials[2]. Among the academic relays, let's mention Bernard Rentier (see Figure 1) a Belgian virologist, former rector of the University of Liège and active promoter of open access policies: *"Very bad news, the (hopefully temporary) closure of the excellent @Covidata.be, exasperated by the lack of transparency on #Covid_19 information in Belgium. Beware: opacity is the main stimulus for conspiracy theorists"*[3]. However, over time, these open data have proven their usefulness, such as the French website covidtracker.fr, developed by Guillaume Rozier (and linked to the popular service ViteMaDose).

In this article, we propose to draw up an inventory of open data on COVID-19 in Belgium, to identify the uses and the obstacles to reuse, and finally to analyze the available data sets. To do so, we rely on a case study (anatomy of the pandemic in Belgium) based on published data. Our article is divided into four parts. The first part is dedicated to the review of the literature and will be followed by the presentation of the methodology. The third part, divided into three sub-parts, is devoted to the presentation of the results. The fourth part is committed to the discussion of the results.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
OpenSym 2021, September 15–17, 2021, Online, Spain
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8500-8/21/09…$15.00
https://doi.org/10.1145/3479986.3479988

[1] Cf. https://www.rtbf.be/info/belgique/detail_quand-le-federal-refuse-d-ouvrir-les-donnees-sur-l-epidemie-de-coronavirus-un-probleme-majeur-qui-retarde-les-scientifiques?id=10466339

[2] Cf. http://margauxdere.be/pourquoi-lopen-data-est-important-pour-vaincre-la-covid-19/ or the Twitter account of the Hainaut deputy Catherine Fonck (former doctor)

[3] Cf. https://twitter.com/bernardrentier/status/1347839421899550721

Figure 1: Relay of a tweet in support of the Belgian site covidata.be

2 STATE OF THE ART

2.1 Open data

The term "open data" refers to "*information that has been made technically and legally available for reuse*" (19). Data release in the public sector is motivated by transparency (e.g., budget analysis) and the belief that data, like software, are "*public goods*" ([3]; [19]). This practice can also facilitate data sharing between public organizations, foster the emergence of new applications useful to citizens (e.g., public transportation), or improve data quality and enrichment through crowdsourcing initiatives ([3]; [8]; [19]; [20]). Huijboom and Van den Broek [15], for their part, distinguish 3 generic strategies: improving democratic control and representation; strengthening policing and law enforcement; and product and service innovation. The authors show different strategic orientations between states and a more or less pronounced focus in different cases. Open data initiatives in the European public sector are promoted by European Directives such as PSI (2003/98/EC) or INSPIRE (2007/2/EC). Several initiatives are already popular in the UK, the US, France or Brazil (e.g. data.gov.uk, thedatahub.org, data.gov, data.gouv.fr, data-publica.com, www.dados.gov.br...) ([3]; [2]). Nikiforova [23] conducted an early analysis of open data policies related to COVID-19. The results showed that 32 out of 52 national open data portals were publishing such open data while Austria, France, Switzerland, and the US were the most responsive in publishing open data within the first two weeks of the pandemic. These open data can in practice feed into open science initiatives [11]. For example, in the spirit of reproducible research, an epidemiological model is described in a publication accompanied by an open source implementation and open data inputs and results.

2.2 Database

Organizations are required to build databases and operate them as part of a decision-making process [36]. In the case of COVID-19, the organizations involved are those involved in the state health response and must make decisions (e.g., confinement decision and configuration of confinement policy based on hospital congestion status and the extent of virus circulation). In this context, the value of the information will depend on the use made of it and will generally decrease over time. This information is actually a representation of the reality in which we are interested. The representations will be built from a set of signals supposed to translate the states of the real world through an information function [36]. These representations are materialized by the data present in the database. These data are subsequently transformed into information through

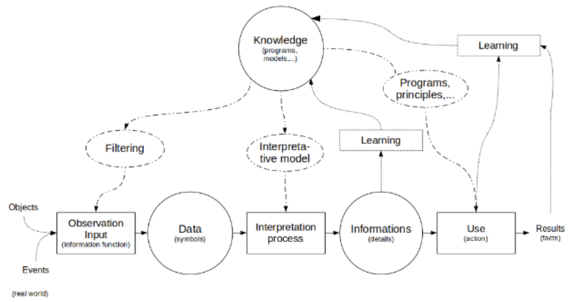

Figure 2: Data valuation process (adapted from Reix et al., 2011)

an interpretation process. This process is based on pre-existing knowledge (in particular the interpretative model used). The use of information leads to results, which are sources of learning and new knowledge, allowing not only the interpretation of data but also their selection during acquisition (filtering).

This process, described by Reix et al. (2011), is shown in Figure 2. It is also described, in a more popular form, as the DIKW pyramid [10]: Data, Information, Knowledge and Wisdom. In this form, data are described as facts, raw materials, that have been accumulated over time "*by people or by machines from observation*"[4]. Information is well-formed data to which meaning has been added. Knowledge involves data, meaning and practice. It is "*a resource for an entity's ability to act effectively*". Wisdom is "*the ability to make optimal use of knowledge to establish and achieve desired goals*". It can be individual (competence) or organizational (capacity). The latter can be related to absorptive capacities [36] or to the dynamic capacities dear to Teece [33].

2.3 Relevance of data

Organizations therefore make decisions based on representations of reality (e.g., management dashboard). Are these representations "*relevant*"? The relevance of representations refers to their use. What is relevant is "*what is appropriate, what is suitable for action*". In other words, data are useful when they lead to individual and collective wisdom. Relevance is therefore "*a quality relative to a user and a context of use*" [36]. Several quality criteria are used to judge the relevance of representations. Reix et al. (2011) cite three main ones:

[4]Translations from references in French are made by the author

completeness, accuracy, and degree of finesse. The quality of completeness refers to the completeness of the representation: all significant changes of state of reality are covered by the representations (first order risk). The accuracy quality refers to the absence of noise: the representations should ideally avoid the second kind of risk related to the taking into account of events caused by random variations due to imperfections of the information function. The degree of finesse concerns the precision of the representation, i.e. its level of detail or its range of variation. In addition, according to Reix et al (2011), there is the punctuality (respect of deadlines), the reliability (confidence in the source), the form (data, drawings, still or animated images...) and the accessibility (search and access methods).

3 METHODOLOGY AND DATA

The Google search engine was used, on the one hand, to identify open data sources related to COVID-19 in Belgium, on the other hand, to identify press articles dealing with open data. In the first case, the queries *open data "covid-19" site:be* and *"open data" "covid-19" site:be* were used. In the second case, the reference news site RTBF Info was targeted using the *"open data" "covid-19" site:rtbf.be* query. Only articles discussing the Belgian open data policy COVID-19 were taken. In particular, RTBF articles simply relaying pandemic figures based on Sciensano open data as part of their information mission were not taken into account. Seven articles developing criticisms on the Belgian open data strategy were thus retained.

The case study was based on the data provided by the governmental reference sites Sciensano and Stabel as well as on the data provided by Google on mobility. The open source software LibreOffice.org Calc (spreadsheet) and R (statistics) were used to process this data.

4 RESULTS: COVID-19 OPEN DATA IN BELGIUM

4.1 Data sources

The unavailability of open data in Belgium was identified at the beginning of the pandemic as an obstacle to the work of scientists[5]. The publication of open data files is the responsibility of Sciensano, a public institution created on February 25, 2018 and the result of the merger between the former "*Centre d'Études et de Recherches Vétérinaires et Agrochimiques*" (CERVA) and the former Scientific Institute of Public Health (ISP).

In practice, the data were published until now in PDF format, which did not allow easy processing in spreadsheet or statistical software. Note that open data are classically published on different portals reflecting the institutional structure of the country. The federal level proposes the portal Data.fgov.be publishing its own data and referencing data from other portals. Other important portals include the portal of the federal statistical institute Statbel, the WalStat portal of the Walloon statistical institute IWEPS and the ODWB portal of the Wallonia-Brussels Federation (FWB).

In practice, the federal portal provides a central view of the characteristics of the datasets made available in open data on its Datasets page. We can see that a large set of formats is supported with a predominance of CSV (textual tabular data), WMS (image format for maps) and JSON (serialization format used by web developers). As for the licenses, they reflect the importance of the referenced datasets: Open Data License Flanders for the Flemish portal, Statbel Open License for Statbel, Etalab Open License for the City of Brussels... As for Sciensano, it publishes its raw data, notably relating to COVID-19[6], on a separate site. The license of the data does not appear clearly on the Epistat site but appears[7] on the FAIR site (Open Data Commons Attribution License). The supported formats are CSV and JSON for the specific data (number of deaths, number of hospitalizations, number of confirmed cases...) and XLS (default format for older versions of Microsoft Excel) for the complete tabular dataset. Although proprietary and supplanted by more recent formats (ODS, XLSX...), the XLS format is still frequently used because it is widely supported as an import or export format for many software programs on the market.

4.2 Reuse of data

These data are important for several reasons. First, they meet communication needs, particularly in the media, in order to explain the characteristics of the pandemic in a more pedagogical way. This is notably the work done by Covidata, an initiative covering a group of researchers proposing analyses and graphics under a CC0 license (CreativeCommons public domain license) accompanied by a set of open source creation scripts published on Github[8] (see Figure 3 for an example of a visualization superimposing the first and second waves with curfew and confinement dates).

Secondly, beyond visualization, they feed into the creation of indicators to better understand the evolution of the pandemic, possibly by combining Sciensano data with data provided by other organizations such as Statbel. We develop an example in section 4.3. There we present a COVID-19 excess mortality index by comparing the proportion of deaths per age group and per region and the respective importance of each age group in the population of each region (see Table 1), which makes it possible to highlight an excess mortality in Wallonia and, above all, Brussels compared to Flanders that cannot be explained by the difference in population density alone.

Thirdly, open data facilitate the development of simulation models that allow a better understanding, and therefore anticipation, of the dynamics of the pandemic at a local level. Nicolas Vandewalle, Professor at the University of Liège, has developed a SEIR model, published on Github[9], which allows to visualize the evolution of the pandemic and to anticipate the saturation of intensive care units, thus providing a tool for decision support (e.g. confinement).

4.3 Case study

The case study proposed here consists of an analysis of the COVID-19 pandemic in Belgium with a specific focus on regional differences. Belgium is indeed a federal state. While health remains an essentially federal competence (e.g. hospitals and social security), on

[5] Cf. https://www.rtbf.be/info/belgique/detail_quand-le-federal-refuse-d-ouvrir-les-donnees-sur-l-epidemie-de-coronavirus-un-probleme-majeur-qui-retarde-les-scientifiques?id=10466339

[6] Cf. https://epistat.wiv-isp.be/covid/

[7] Cf. https://fair.healthdata.be/dataset/9c20457b-e66d-4377-ab8e-d48f6dc2b034

[8] Cf. https://github.com/pschaus/covidbe-opendata

[9] Cf..https://github.com/glouppe/covid19be

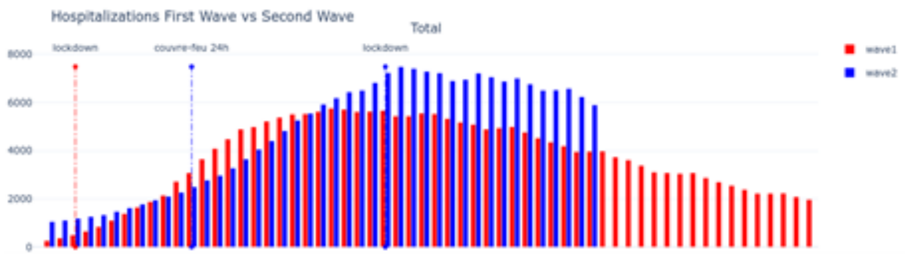

Figure 3: Covidata.be visualization of the number of hospitalizations in open data

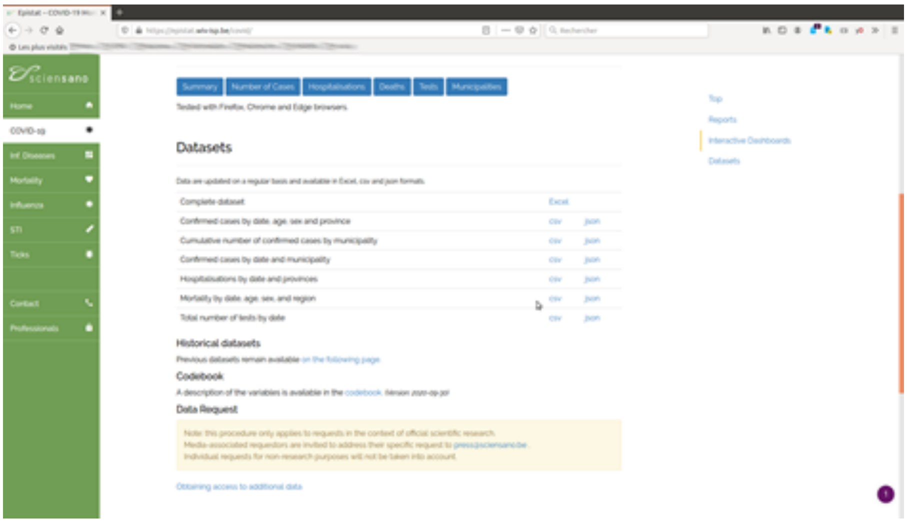

Figure 4: COVID-19 Sciensano open data portal

Table 1: Mortality by age group in Belgium

Age	Deaths (#)	Deaths (%)	Cumulated
0-24	4	0.03 %	99.85 %
25-44	57	0.42 %	99.82 %
45-64	729	5.38 %	99.40 %
65-74	1626	11.99 %	94.02 %
75-84	3999	29.49 %	82.03 %
85+	7125	52.54 %	52.54 %
NA	21	0.15 %	

the one hand, the overall response to the pandemic depends on other federated entities (e.g. community, for testing in educational institutions, and region for health response in nursing homes), and on the other hand, the regions present distinct social and economic realities.

In practice, Sciensano provides its data in the form of text files in CSV format or workbooks in Microsoft Excel format (see Figure 4). A file containing all the data sets is provided in the latter format. This is the file we work on (dated November 11, 2020). These datasets are provided with documentation on the semantics of the data[10]. The data are generally structured along different dimensions: region, province, city, age (range), gender... The data are of course aggregated, without personal data. The values provided concern the number of deaths, hospitalizations, patients in intensive care and tests.

These files are therefore designed to be easily processed from specific software (e.g. reading CSV files from a Python script) or

[10]Cf. https://epistat.sciensano.be/COVID19BE_codebook.pdf

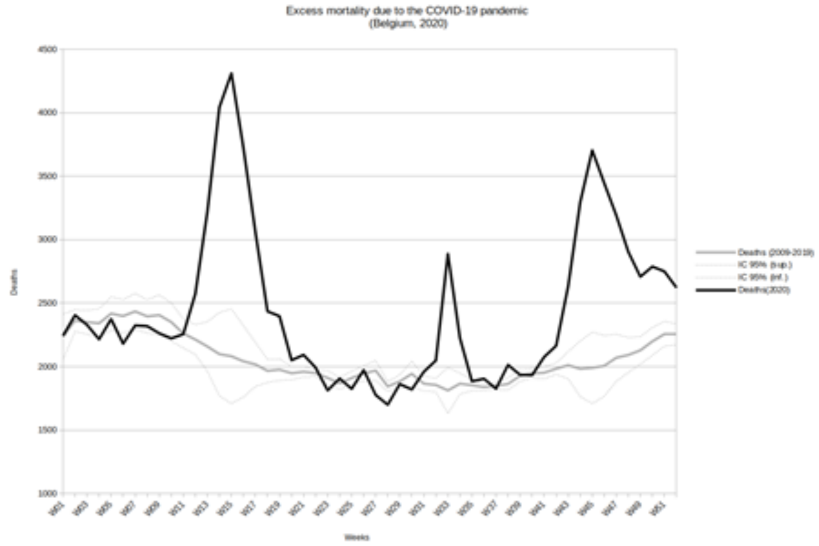

Figure 5: Excess mortality related to the COVID-19 pandemic in Belgium

Table 2: Excess mortality by age group and region

Age	Flanders	Wallonia	Brussels
25-44	-43,45 %	19,42 %	137,26 %
45-64	-45,96 %	38,76 %	149,15 %
65-74	-36,63 %	33,36 %	156,11 %
75-84	-26,48 %	29,64 %	114,69 %
85+	-20,28 %	24,04 %	69,52 %

from standard spreadsheet software using pivot tables or standard functions (NB.SI, SUM.SI...). An open source spreadsheet program such as LibreOffice.org Calc allows for example to easily calculate mortality by age group as well as cumulative mortality starting from the oldest population (see Table 1).

To visualize the existence of excess mortality in the year 2020 during which the pandemic occurred, one has to use the Statbel[11] data provided for the period 2009-2021, which allows to visualize the excess mortality in 2020 during successive waves (cf. Figure 5). These data are also exploited by Sciensano in the Be-MOMO[12] project, for the analysis of excess mortality due to COVID-19 and the validation of the methodology for calculating mortality due to COVID-19 (31) .

The data provided allow the same calculation to be performed by region so as to determine mortality at the national and regional levels (Flanders, Brussels, and Wallonia). However, the comparison is not so simple. Indeed, these regions present distinct realities, notably demographic[13], and we have just seen (Table 1) that the older populations paid the major part of the price in this pandemic. Thanks to the official Statbel website, it is possible to know the population part (%) by age group and by region. From this, it is possible to calculate the excess mortality per age group by dividing the share of COVID-19 deaths and the share of living persons by age group and region (see Table 2). This calculation (i.e. the ratio, for a given province and age group, between the percentage of COVID-19 deaths and the percentage of this sub-population in the general population) shows a high under-mortality in Flanders, given the greater age of the population in the north of the country.

Several factors could explain this excess mortality, such as population density and poverty level. The standard of living, in terms of per capita income, is indeed higher in Flanders[14]. Wallonia, for example, has a high level of unemployment[15] and a lower average per capita GDP (Gross Domestic Product) in its former industrial

[11]Cf. https://statbel.fgov.be/fr/open-data/nombre-de-deces-par-jour-sexe-arrondissement-age
[12]Cf. https://epistat.wiv-isp.be/momo/
[13]Cf. https://statbel.fgov.be/sites/default/files/images/in%20de%20kijker/Chiffrescles_2019_r.pdf, (see page 14 for a breakdown by age group)
[14]Cf. https://www.iweps.be/indicateur-statistique/revenus-menages-habitant/ for a comparison between regions
[15]Cf. https://www.iweps.be/indicateur-statistique/taux-de-chomage-administratif-15-a-64-ans/

Table 3: Linear regression in R

```
Residuals:
     Min        1Q    Median        3Q       Max
-0.189992 -0.119358 -0.006006  0.107728  0.242161
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.961e+00  3.999e-01   7.403 4.09e-05 ***
RHab        -1.600e-04  2.111e-05  -7.581 3.39e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.1453 on 9 degrees of freedom
Multiple R-squared:  0.8646,  Adjusted R-squared:  0.8496
F-statistic: 57.48 on 1 and 9 DF,  p-value: 3.391e-05
```

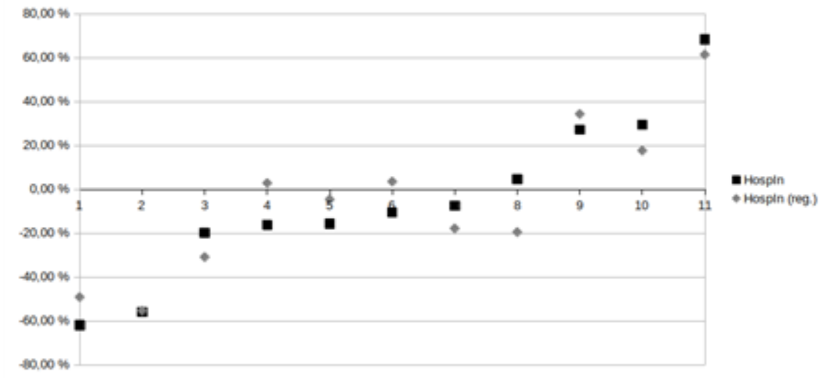

Figure 6: Hospitalization figures (actual vs regression)

area (Mons-Charleroi-Liège axis) [5], while Brussels, although generating a per capita GDP well above the European average [5], is fed by Walloon and Flemish commuters who work but do not reside in Brussels[16], and has a fairly poor population[17] on average, particularly as a result of immigration [25]. Several useful data can be retrieved in this respect: GDP per capita (e.g. IWEPS and Statbel), income per region (Statbel) or specific poverty indicators (e.g. AROPE provided by Statbel). The correlation between per capita income and hospitalization rate, by province, appears to be higher (than that with AROPE or density: -0.930 vs. 0.823 vs. 0.637), so per capita income is used to run a linear regression to predict the hospitalization rate. The latter is used instead of mortality because, at the time of this case (November 2020), mortality by province was not provided, unlike hospitalization figures. The regression model here consists of predicting the excess hospitalization (share of hospitalizations divided by the share of the province's population in the total population) by per capita income. The model provides a satisfactory estimate (see Figure 6): the variable "RHab" is significant and the model has a coefficient of determination R^2 equal to 0.8496 (see Table 3).

The data provided allow for the exploration of other hypotheses (e.g., testing policy; the latter is found to be fairly uniform across the country, for example, regardless of the density of the provinces considered). Other open data can be used, including those provided by private providers such as Google (see Google's "COVID-19 Community Mobility Reports" available for download in CSV format). The latter allow for the analysis of the evolution of mobility related to leisure activities (e.g., increased mobility to the coast during the spring or summer vacations, including the July 21 festivities; see Figure 7) or to work activities (e.g., effectiveness of the confinement and reduction of commuting in Brussels).

4.4 Criticism of the open data strategy

Sciensano's open data policy has been criticized since the beginning of the crisis by prominent French-speaking Belgian personalities such as Bernard Rentier, university pro-rector and promoter of

[16]Cf. https://ibsa.brussels/sites/default/files/publication/documents/FOR_HermReg_2020_12182_F.pdf (page 52 et following)
[17]Cf. https://statbel.fgov.be/fr/themes/menages/revenus-fiscaux

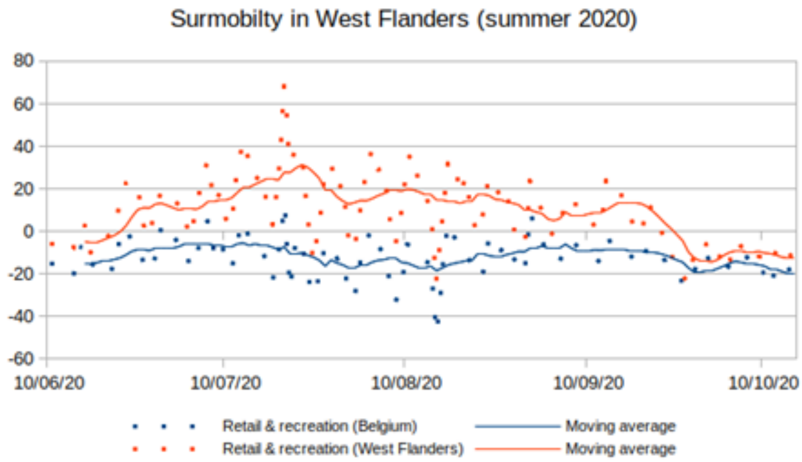

Figure 7: Overmobility at the Belgian coast (summer 2020)

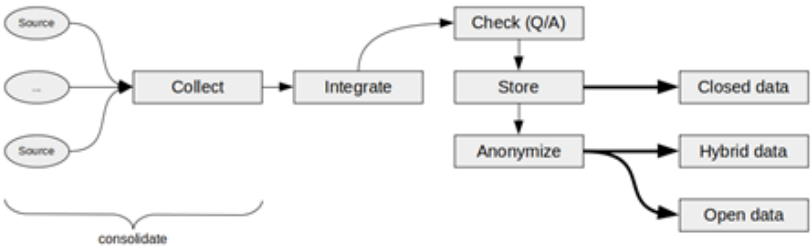

Figure 8: Open data publication process

open science, Pierre Schauss, professor of computer science at the initiative of covidata.be, or Marius Gilbert, epidemiologist: "*The scientific community does not currently have access to the raw data concerning covid-19. This is a major problem that delays us in the answers that can be given to this epidemic. We urgently need to move to open data as Italy and France are doing*"[18]. Basically, Sciensano is mandated to provide data to public authorities.

These early complaints led to the publication of data in open data by the end of March 2020 (the data was then communicated during press conferences or within PDF files that were difficult to process in an automated way). However, tensions with scientists continued until the covidata.be website was temporarily closed in January 2021 in protest against the increasing share of data released outside of open data repositories[19] (a form allows access to sensitive data within three weeks). As Simon Dellicour, an epidemiologist, put it: "*the fact that in Belgium a very small part of the scientific community has access to all epidemiological data is unjustifiable and non-strategic in terms of missed analytical opportunities*" (idem).

As another example, vaccination data were not released until the end of April 2021[20]. In addition to scientists (e.g., studying the effect of returning from vacation), these data are also of interest to journalists for their news and dashboards on the evolution of the pandemic (e.g., explaining the reasons for confinement measures to the population).

Sciensano, on the other hand, points to the workload required to make the collected data available in open data, citing a shortage of 45 people to handle the additional workflow[21]. The difficulty is found at two levels: at the level of the organization itself and at the level of the organization. On the one hand, the collection of data from different sources entails a lot of work to consolidate a network of data collection, to integrate these data and to ensure quality assurance, in particular to validate the published data (cf. Figure 8): "*It would not have been responsible to throw the data away*" (idem).

[18]Cf. https://www.rtbf.be/info/belgique/detail_quand-le-federal-refuse-d-ouvrir-les-donnees-sur-l-epidemie-de-coronavirus-un-probleme-majeur-qui-retarde-les-scientifiques?id=10466339

[19]Cf. https://www.rtbf.be/info/societe/detail_open-data-et-sciensano-manque-de-coherence-entre-le-discours-officiel-et-la-realite?id=10668706

[20]Cf. https://www.rtbf.be/info/societe/detail_campagne-de-vaccination-contre-le-covid-19-le-point-en-chiffres-et-graphiques-ce-dimanche-16-mai?id=10730805

[21]Cf. https://www.rtbf.be/info/belgique/detail_sciensano-notre-premiere-preoccupation-n-etait-pas-de-faire-un-beau-site-internet-en-open-data?id=10509727

Table 4: COVID-19 datasets in open data and by type of source

Government open data(examples)		Private open data(examples)	Scientific open data(examples)
Federal	Sciensano (Epistat, FAIR Healthdata), Statbel[25]	COVID-19 Community Mobility Reports (Google)	Covidata.be (open science: open data + open source)
Region	AVIQ (vaccination statistics in Wallonia[26]		
Community	Open Data Wallonie Bruxelles		
Province	-		
Municipality	-		

On the other hand, data published in open data must be anonymized (e.g., aggregated) and its irreversibility must be guaranteed[22].

5 DISCUSSION

5.1 Characterization of the data sets

The open data identified in this work (see Table 4) are of different natures: either governmental data or private data (made available following the pandemic in the case of Google), or academic data (sometimes in an open science context mobilizing data but also software). The governmental data reflect the Belgian institutional structure but the data specific to the COVID-19 pandemic are fortunately centralized by a unique organization (Sciensano) resulting in frequent redirections to Sciensano from other data sources.

On the licensing side, there are several cases available. In particular, in the case of Sciensano, some of the data is provided on the site under a public domain license while other data is provided upon request: "*Since 31 March 2020, Sciensano has made daily updated data publicly available. Obtaining access to additional data other than these public (open) data is subject to at least: Compliance with the General Data Protection Regulation (GDPR) and the Belgian Law of 30 July 2018 on the protection of individuals with regard to the processing of personal data (including combined datasets consisting partly of personal data and partly of non-personal data); Obtaining authorization of the Information Security Committee, if health data are concerned*".

5.2 Relevance of the representations

Published data raise questions about their relevance (cf. 3.2) in the sense of Reix et al. (2011). The following is a set of potential problems observed following the use of these data or the monitoring carried out on the subject of the use of COVID-19 open data.

Accessibility - Some data are not available in open data (e.g., delays in the availability of vaccination figures in Belgium (and France), added by Sciensano on April 27, 2020. This same observation could be made for the dissemination of wastewater data (cf. Obépine network for example in France). Moreover, some data have been published, but in PDF format, which has handicapped the automated processing on these data.

Accuracy - Test data are affected by the testing techniques that are or are not accounted for (e.g. RT-PCR, antigenic and salivary) in the statistics, which makes comparisons between states or over long periods of time (change of accounting) more complex. In practice, in Belgium, the calculation of the number of cases or the number of tests follows specific rules that Sciensano documents[25]. For example, before March 15, 2020, confirmed cases were in practice, and due to the shortage of tests, possible cases; thereafter, they were cases confirmed by a molecular test (i.e. PCR or rapid antigen test)[26]. Salivary tests, although considered interesting by Sciensano, are therefore not counted[27]. Similarly, the way in which COVID-19 deaths are counted may vary by country and, as in Belgium, may deviate from WHO recommendations [22].

Reliability - Positivity figures should be taken with caution when shortages occur [24]. For example, testing may be prioritized in at-risk or highly symptomatic populations, resulting in higher positivity. Catch-up effects may also occur in the publication of data, leading to upward or downward discontinuities that are not attributable to the observed phenomenon.

Completeness - The effects of confinement measures (e.g. mental health; see [26], on this topic of mental health) are not included in the statistics provided. In addition, data on the use of proximity tracing applications are, as in other countries [34], not public (Belgium).

Fineness - Data are not always provided with the expected granularity (e.g. data by region and not by province in some Sciensano files) but this can be improved over time.

Timeliness - Data are sometimes provided with a delay (e.g. weekend) that disturbs visualization or prediction model results. This problem also exists in Belgium for the monitoring of variants[28].

Punctuality - Sciensano data for Saturday, Sunday and Monday are provided on Tuesday[29].

Form - The processing of open data leads to the creation of new open data as well as to visualizations that facilitate communication around the pandemic (e.g. covidata.be). Graphics can pose different problems (e.g. scale).

[22]Cf. https://www.rtbf.be/info/article/detail_de-sciensano-aux-sites-d-infos-en-passant-par-sydney-comment-les-chiffres-du-covid-arrivent-jusqu-a-vous?id=10717592
[23]Cf. https://statbel.fgov.be/fr/covid-19-donnees-statbel
[24]Cf. https://www.jemevaccine.be/centre-d-informations/un-portail-opendata-vaccination-est-cree-avec-les-chiffres-de-la-vaccination-par-commune-en-wallonie/)
[25]Cf. https://epistat.sciensano.be/COVID19BE_codebook.pdf and https://covid-19.sciensano.be/sites/default/files/Covid19/COVID-19_FAQ_FR_final.pdf
[26]Cf. https://covid-19.sciensano.be/sites/default/files/Covid19/20201012_Advice%20RAG_testing_update%20October_Fr.pdf
[27]Cf. https://covid-19.sciensano.be/sites/default/files/Covid19/20201012_Advice%20RAG_testing_update%20October_Fr.pdf
[28]This information has for example been added to the open data downloadable in France, cf. https://www.data.gouv.fr/fr/datasets/donnees-de-laboratoires-pour-le-depistage-indicateurs-sur-les-variants/
[29]Cf. https://www.rtbf.be/info/societe/detail_coronavirus-en-belgique-sciensano-ne-publiera-plus-de-bilan-du-coronavirus-les-week-ends-et-le-lundi?id=10525696

Figure 9: Pandemic response and knowledge management

5.3 Knowledge management

As emphasized by Rowe et al. [29], the effectiveness of the organized response to the COVID-19 pandemic depends on the management of knowledge about the virus to allow for the adequacy and adaptation of measures taken. Thus, the mechanisms of transmission of the virus are the subject of a continuous work of updating of knowledge. In contrast to SARS [13], the importance of presymptomatic persons (persons infected without symptoms during the incubation period but symptomatic beyond that period) in the transmission of SARS-CoV-2, which accounts for more than forty percent of infections, has rapidly emerged ([13]; [32]; [18]). On the contrary, the role of asymptomatic persons (infected persons with no symptoms), about fifteen percent of the patients [14], in the chain of infection may have been downplayed over time ([21]; [4]). As for the modes of transmission of the virus, while there is a general consensus on direct or indirect contact (fomite) and droplets, the possible contribution of aerosols (particles between 5 and 20 μm in size) to airborne transmission was still being debated in March 2021 ([12]; [18]), although this scientific debate does not negate the importance of improving ventilation in enclosed spaces, for example [1].

This evolving nature of knowledge in the face of a new phenomenon implies continuous monitoring and consolidation of knowledge to enable decision-making in the various dimensions of action in the face of the pandemic [34] (see Figure 9). This characteristic also influences the reuse of open data. For example, in the case of models, their execution assumes the availability of relevant data to feed them but also of up-to-date knowledge to configure them, knowledge itself sometimes resulting from a process of progressive data refinement (cf. DIKW model; [10]). Thus, for example, in a model taking into account the diffusion of the virus, the share of contaminations due to presymtomatics must be known and adjusted as soon as the scientific consensus is established or evolves. Similarly, the rate of vaccination, the importance of herd immunity or the importance of immune escape [16] will influence the work of modelers, and presupposes both up-to-date knowledge and the availability of quality raw data.

5.4 Comparison with France

The French situation with regard to the processing of COVID-19 data is also beginning to be documented. Ronai [27] attributes the French delays in publishing complete mortality data to technical and organizational problems. The author distinguishes between a statistically oriented information system, with an administrative channel (INSERM) and a health channel (INSEE), and a real time information system (SI-VIC). The latter, initially deployed to record the victims of the Paris attacks of November 2015, was eventually used for real-time monitoring of COVID-19-related mortality. Initially, it only counted deaths that occurred in hospital and excluded deaths at home or in retirement homes (i.e. 44% of COVID-19 deaths[30]. The latter could finally be counted via a fourth system (Voozanoo) dedicated to EHPADs. The data are therefore partitioned between different systems and must be consolidated before publication. Under pressure from groups of data scientists, such as OpenCOVID19, SantéPubliqueFrance will gradually open up its data sets from March 18, 2020 ([7]; [27]). This will allow several high-profile initiatives to be launched (e.g. CovidTracker). For example, statistics on hospitalizations according to vaccination status were only published by DREES in August 2021 (the breakdown of these data by age group is announced for later), whereas the vaccination campaign started in France on 27 December 2020[31]. They thus required the merging of data from SI-VIC (hospitalization), SI-DEP (screening) and VAC-SI (vaccination)[32]. No matter how much professionals may want to, the production of real-time data sets and dashboards for evidence-based decision-making is not immediate.

6 CONCLUSION

In this exploratory research, we analyzed the open data policy implemented in Belgium. We described the available datasets as well as the applications to which these datasets had led. Based on a case study and a review of press articles, we showed the current limitations of Belgian open data in terms of the relevance of the published data.

This research is currently limited to one country (Belgium). It should be extended to other neighboring countries, which would allow a comparison of publication policies, for example between Belgium and France, which is close to Belgium and is rather among the good students in terms of open data [23], despite the criticisms expressed by the most active reusers[33].

[30]Cf. https://www.lemonde.fr/les-decodeurs/article/2020/12/03/les-residents-d-ehpad-representent-44-des-morts-du-covid-19_6062084_4355770.html)

[31]Cf. https://www.santepubliquefrance.fr/dossiers/coronavirus-covid-19/vaccination-contre-la-covid-19

[32]Cf. https://data.drees.solidarites-sante.gouv.fr/explore/dataset/covid-19-resultats-issus-des-appariements-entre-si-vic-si-dep-et-vac-si/information/?disjunctive.vac_statut

[33]Frenchman Guillaume Rozier, developer of CovidTracker and promoter of open data, thus regularly communicates about the slow pace of data communication (e.g. variant figures). See https://www.lamontagne.fr/paris-75000/actualites/pour-guillaume-rozier-de-covidtracker-l-open-data-permet-de-lutter-contre-la-defiance-et-les-complotistes_13925014/ for an overview.

The understanding of the seemingly more important obstacles to the publication of open data in Belgium would merit further investigation in order to distinguish, and weigh up, cultural, legal, organizational and technical causes. This research could be done through a set of semi-structured interviews in organizations active in health, including IT development structures (e.g. SMALS) and institutions dedicated to public health (e.g. Sciensano).

REFERENCES

[1] Bhagat, R. K., Wykes, M. D., Dalziel, S. B., & Linden, P. F. (2020). Effects of ventilation on the indoor spread of COVID-19. Journal of Fluid Mechanics, 903.
[2] Breitman, K., Salas, P., Casanova, M. A., Saraiva, D., Gama, V., Viterbo Filho, J., & Chaves, M. (2012). Open government data in Brazil. IEEE Intelligent Systems, 27(3), 45-49.
[3] Brugière, A., & Népote, C. (2011). Guide pratique de l'ouverture des données publiques territoriales. FING.
[4] Cao, S., Gan, Y., Wang, C., Bachmann, M., Wei, S., Gong, J., & Lu, Z. (2020). Post-lockdown SARS-CoV-2 nucleic acid screening in nearly ten million residents of Wuhan, China. Nature communications, 11(1), 1-7.
[5] Capron, H. (2009). La compétitivité des régions. Reflets et perspectives de la vie économique, 48(1), 115-136.
[6] Chang, S., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D., & Leskovec, J. (2020). Mobility network models of COVID-19 explain inequities and inform reopening. Nature, 1-8.
[7] Chignard, S. (2021). L'open data de crise : entre mobilisation citoyenne et communication gouvernementale. Enjeux numériques. Les Annales des Mines, n°14, juin 2021, pp. 73-77.
[8] Chignard, S. (2012). Open Data. FYP Editions, France.
[9] Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. The Lancet infectious diseases, 20(5), 533-534.
[10] Ermine, J. L., Moradi, M., & Brunel, S. (2012). Une chaîne de valeur de la connaissance. Management international/International Management/Gestion Internacional, 16, 29-40.
[11] Fecher, B., & Friesike, S. (2014). Open science: one term, five schools of thought. Opening science, 17-47.
[12] Greenhalgh, T., Jimenez, J. L., Prather, K. A., Tufekci, Z., Fisman, D., & Schooley, R. (2021). Ten scientific reasons in support of airborne transmission of SARS-CoV-2. The Lancet.
[13] He, X., Lau, E. H., Wu, P., Deng, X., Wang, J., Hao, X., & Leung, G. M. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. Nature medicine, 26(5), 672-675.
[14] He, J., Guo, Y., Mao, R., & Zhang, J. (2020). Proportion of asymptomatic coronavirus disease 2019: A systematic review and meta-analysis. Journal of medical virology.
[15] Huijboom, N., & Van den Broek, T. (2011). Open data: an international comparison of strategies. European journal of ePractice, 12(1), 4-16.
[16] Kupferschmidt, K. (2021). New mutations raise specter of 'immune escape', Science, 22 Jan 2021 : 329-330.
[17] Lau, H., Khosrawipour, V., Kocbach, P., Mikolajczyk, A., Schubert, J., Bania, J., & Khosrawipour, T. (2020). The positive impact of lockdown in Wuhan on containing the COVID-19 outbreak in China. Journal of travel medicine, 27(3).
[18] Lau, E. H., & Leung, G. M. (2020). Reply to: Is presymptomatic spread a major contributor to COVID-19 transmission?. Nature Medicine, 26(10), 1534-1535.
[19] Lindman, J., & Tammisto, Y. (2011, October). Open Source and Open Data: Business perspectives from the frontline. In IFIP International Conference on Open Source Systems (pp. 330-333). Springer, Berlin, Heidelberg.
[20] Lobre, K. (2012). L'Open Data en 2012 : panorama des risques. AIM 2012, 21-23 mai 2012.
[21] Madewell, Z. J., Yang, Y., Longini, I. M., Halloran, M. E., & Dean, N. E. (2020). Household Transmission of SARS-CoV-2: A Systematic Review and Meta-analysis. JAMA network open, 3(12).
[22] Molenberghs, G., Faes, C., Verbeeck, J., Deboosere, P., Abrams, S., Willem, L., & Hens, N. (2020). Belgian COVID-19 Mortality, Excess Deaths, Number of Deaths per Million, and Infection Fatality Rates (9 March-28 June 2020). medRxiv.
[23] 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.
[24] Omori, R., Mizumoto, K., & Chowell, G. (2020). Changes in testing rates could mask the novel coronavirus disease (COVID-19) growth rate. International Journal of Infectious Diseases, 94, 116-118.
[25] Perrin, N., & Van Robaeys, B. (2006). La pauvreté chez les personnes d'origine étrangère chiffrée. Fondation Roi Baudouin.
[26] Pfefferbaum, B., & North, C. S. (2020). Mental health and the Covid-19 pandemic. New England Journal of Medicine, 383(6), 510-512.
[27] Ronai, M. (2021), La construction d'un système d'information épidémiologique, Annales des Mines, n°14, juin 2021.
[28] Rowe, F., Fallery, B., Reix, R., & Kalika, M. (2011). Systèmes d'information et management des organisations. Vuibert.>
[29] Rowe, F., Ngwenyama, O., & Richet, J. L. (2020). Contact-tracing apps and alienation in the age of COVID-19. European Journal of Information Systems, 29(5), 545-562.
[30] Saglietto, A., D'Ascenzo, F., Zoccai, G. B., & De Ferrari, G. M. (2020). COVID-19 in Europe: the Italian lesson. Lancet, 395(10230), 1110-1111.
[31] Sierra, N. B., Bossuyt, N., Braeye, T., Leroy, M., Moyersoen, I., Peeters, I., & Renard, F. (2020). All-cause mortality supports the COVID-19 mortality in Belgium and comparison with major fatal events of the last century. Archives of Public Health, 78(1), 1-8.
[32] Slifka, M. K., & Gao, L. (2020). Is presymptomatic spread a major contributor to COVID-19 transmission?. Nature Medicine, 26(10), 1531-1533.
[33] Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. Strategic Management Journal, 18(7), 509-533.
[34] Viseur, R. (2021). Stratégies numériques de lutte contre la pandémie de la COVID-19 : quelle place pour les applications de contact tracing ? Congrès InforSID, Dijon (France).
[35] Xu, B., Gutierrez, B., Mekaru, S. *et al.* (2020). Epidemiological data from the COVID-19 outbreak, real-time case information. Sci Data 7, 106.
[36] Zahra, S. A., & George, G. (2002). Absorptive capacity: A review, reconceptualization, and extension. Academy of Management Review, 27(2), 185-203.
[37] Reix, R., Fallery, B., Kalika, M., & Rowe, F. (2011). Systèmes d'information et management des organisations (6e édition.). Paris: Vuibert.