doi:10.3233/SHTI210816

Beyond the FAIRness of COVID-19 Data: What about Quality?

Fabrizio PECORARO[1] and Daniela LUZI
Institute for Research on Population and Social Policies, National Research Council, Rome, Italy

Abstract. Different datasets have been deployed at national level to share data on COVID-19 already at the beginning of the epidemic spread in early 2020. They distribute daily updated information aggregated at local, gender and age levels. To facilitate the reuse of such data, FAIR principles should be applied to optimally find, access, understand and exchange them, to define intra- and inter-country analyses for different purposes, such as statistical. However, another aspect to be considered when analyzing these datasets is data quality. In this paper we link these two perspectives to analyze to what extent datasets published by national institutions to monitor diffusion of COVID-19 are reusable for scientific purposes, such as tracing the spread of the virus.

Keywords. FAIR, data quality, COVID-19, institutional datasets, data reusability

1. Introduction

Already from the beginning of the COVID-19 pandemic in March 2020 national and international authorities started to develop and update datasets to provide data to researchers, journalists, health care providers as well as public opinion. This data became one of the most important sources of information, commonly daily updated, to be analysed by scientists to investigate this epidemic period. Data is examined by the research community not only to monitor the COVID-19 diffusion across countries and localities for research purposes, but also to gain insights and propose better containment measures and policies. To facilitate the comparability and reuse of this data, one of the first target is to make these datasets compliant with the FAIR (Findability, Accessibility, Interoperability, and Reusability) principles [1]. These principles are gaining consensus within scientific communities, with different initiatives carried out in the healthcare domain [2] at national and international level with the aim of promoting their adoption and implementation when defining and sharing research data. Despite compliance with the FAIR principles is mainly met to research results, such as clinical trials or human genomics, in this paper we pose the attention on datasets published by national institutions to report aggregate data on the diffusion of COVID-19. Furthermore, even if the compliance with the FAIR principles may be considered as a proxy for data quality assessment, they do not, in themselves, cover the crucial aspects of intrinsic data quality. However, to establish credibility

[1] Corresponding Author, Fabrizio Pecoraro, IRPPS-CNR, via Palestro 32, 00185 Rome, Italy; E-mail: f.pecoraro@irpps.cnr.it

studies that use healthcare data are increasingly expected to demonstrate that the quality of data is adequate to support research conclusions [3]. This is particularly true considering COVID-19 surveillance data that represents an essential tool to monitor trends in the epidemics, to conduct risk assessments and to timely guide preparedness and response measures [4]. For these reasons, aim of this paper is to capture the level of FAIRness of the above-mentioned institutional datasets also under the lens of the data quality model proposed by the ISO 25012 [5] which is used to define data quality requirements guiding software development.

2. Materials and Methods

COVID-19 institutional datasets available at national level in six European countries (Belgium, France, Germany, Italy and UK) were included in the analysis. They were identified carrying out a literature review in the LitCovid [6] portal that tracks COVID-19 related articles in PubMed. In particular, we concentrated on the Epidemic Forecasting section to identify datasets adopted to model the spread of COVID-19 focusing on at least one of the above-mentioned countries. Data availability statement of each paper has been analysed to extract the source of information applied to perform the analysis. Results of this review are updated at the end of June 2021.

The extracted datasets have been firstly analysed under the FAIRness perspective checking their compliance to the 15 sub-principles reported in [1]. Considering data quality, different assessment methods and models have been proposed in the literature [7] most of them defined in specific health context (e.g. prevention) or focusing on a specific disease (e.g. cancer). This perspective differentiation led authors to adopt different data quality characteristics depending on relevant points of view. In this paper we adopted the data quality model reported in ISO 25012 [5] which is widely used in different domains both at industrial and scientific levels. This standard is based on 15 characteristics classified into two categories: 1) *inherent data quality* that refers to the degree to which data quality characteristics have intrinsic potential to satisfy implicit data needs and 2) *system-dependent data quality* that refers to the degree to which data quality is achieved and preserved through an information system and is dependent on the specific technological context in which the data is used. In this paper we focus the attention on the inherent data quality characteristics.

3. Results

3.1. Analysis of national datasets on COVID-19

Among the 1700 papers published within the Epidemic Forecasting section of the LitCovid platform, 338 reported information on at least one of the six countries involved in this analysis. Almost three-quarters of them (N = 256) were excluded from the analysis as they are based on datasets published by international bodies (e.g. WHO) or adopted data collected specific studies (e.g. surveys, hospital). Table 1 shows the list of datasets adopted in the 82 remaining papers which also makes references to the institutions that curate them.

Table 1. Source of institutional datasets reported at national level

Country	Publisher	Source / Dataset
Belgium	Sciensano	https://hepistat.wiv-isp.be/Covid/
France	Public Health System	https://www.data.gouv.fr/fr/pages/donnees-coronavirus
Germany	Robert Koch Institute	https://npgeo-corona-npgeo-de.hub.arcgis.com/ https://github.com/jgehrcke/covid-19-germany-gae
Italy	Civil Protection Department	https://github.com/pcm-dpc/COVID-19
Spain	Carlos III Health Institute	https://cnecovid.isciii.es/covid19/; https://github.com/datadista/datasets/tree/master/COVID%2019
UK	Public Health England	https://coronavirus.data.gov.uk/

3.2. Analysis of FAIR principles

Table 2 shows the level of compliance of each dataset to the main FAIR principles. Considering the presentation of data, all countries defined a specific section of the institutional website to describe which data are exposed. Among them, Italy, Germany and Spain adopt the GitHub service that allows the download of CSV and JSON files directly or through the adoption of the GitHub REST API. Similarly, UK and Germany provide data with self-developed API that can also be used to download data in CSV or JSON formats. This presentation of data not only simplify the accessibility of datasets, but also ensures their findability given the permanent link through which researchers can access data routinely. Conversely, data on France and Belgium can be accessed only by downloading CSV files reported in the relevant web pages. In this case the unique identifier as well as its stability is not easily verifiable.

Table 2. Assessment of the FAIR principles in each national institutional dataset

	Belgium	France	Germany	Italy	Spain	UK
Findable						
F1. Unique ID	HTML	HTML	API	GitHub	GitHub	API
F2 & F3. Metadata richness & ID	Limited in PDF (English)	Limited in CSV (English)	Limited in Web pages (German)	Limited in Web pages (English)	Limited in Web pages (English)	Limited in Web pages
F4. Metadata	No	No	No	No	No	No
Accessible						
A1. Retrievability	File	File	API	API	API	API
A1.1. Protocol	CSV	CSV	API	Github	Github	API
A1.2. Auth	N/A	N/A	N/A	N/A	N/A	N/A
A2. Metadata	N/A	N/A	N/A	N/A	N/A	N/A
Interoperable						
I1. Language	No	No	No	No	No	No
I3. Vocabulary	No	No	No	No	No	No
I4. Reference	No	No	No	No	No	No
Reusable						
R1. Accurate	No	No	No	No	No	No
R1.1. License	Open data	Open data	Open data	Open data	Open data	Open data
R1.2. Origin	Not clear	Not clear	Not clear	Partial	Not clear	Not clear
R1.3. Standard	No	No	No	No	No	No

Considering metadata all countries provide a limited set of descriptive information, such as description and data type, along with examples describing them. Moreover, in all countries the association between a metadata file and the dataset is not explicit or even not reported. In particular, Belgium reports a codebook in a PDF file written in

English, while Germany, Italy, Spain and UK report metadata and description of indicators in specific web pages of the dataset website. France is the only country that provides a set of CSV files associated with each CSV data file reporting metadata and information about relevant indicators. However, the association between data and metadata files is not straightforward with no cross references in the documentation. All countries provide access to both data and metadata with no authentication or authorization procedures needed.

Looking at the interoperability principles, the absence of controlled vocabularies, ontologies, thesauri as well as of a data model make the integration of data and the performance of a cross-country analysis hard to be accomplished. Moreover, even if all countries, except Germany, report the description of indicators also in English, variables are generally instantiated using the original language considering both the name and the value of the indicator. The reuse of data for statistical purposes is also affected by the absence of a detailed description of the workflow that led to the collection of data. In particular, data flow and provenance of data are not sufficiently reported in each website, this is mainly critical in regional-based countries where information are daily transmitted by each region to national authorities. Lack of standardized collection of data have been reported in Italy [8] as well in Spain [9] where, each regions might count case numbers and tests with different criteria. Within the reuse of data all countries release data under the Creative Commons rules.

3.3. Analysis of quality characteristics

Considering *credibility* and *traceability* the data flow adopted to collect, elaborate and diffuse data is not reported by the analyzed countries with the exception of Italy, where the data flow is partially descripted leaving out information on data collection time periods at local level and their submission to the relevant region. The feature of *currentness* and in particular data *timeliness* represents one of the positive data quality aspects of COVID-19 datasets. Data are mainly daily updated in all countries at local and national level. On the contrary, datasets lack of data *understandability* as all countries report both the name and the values of each variable in their own originated language making it necessary to translate them before integration. Also the absence of the formulas that clearly describe how each indicator is computed makes the comparability of data particularly complex. Moreover, the level of data *disaggregation* is an important feature to be considered as it allows to compare data across countries and to provide a coherent analysis at European level. With the exception of Italy, the other countries analyzed provide data distributed by gender and age ranges.

4. Discussion and Conclusions

The paper presents an analysis of the FAIRness level of datasets distributed by national authorities to map the spread of COVID-19 in six European countries. Moreover, FAIR principles have been conceptually linked with ISO 25012 considering in particular the characteristics of the *inherent data quality*. This was done to explore whether the minimum set of data description identified by the high level, disciplinary-independent FAIR principles cover the main quality features of data. This extended analysis is particularly important considering the crucial role played by the diffusion of COVID-19 analyses on which researchers and policy makers have relied to face pandemic.

Considering FAIR principles, differences across datasets have been detected in their accessibility and findability. The adoption of GitHub services or customized APIs facilitates the access to data and metadata improving their retrievability thanks to standardised, open and universally implementable communications protocols. Moreover, this solution simplifies the assignment of global unique and persistent identifiers to both data and metadata. Conversely, considering the interoperability and reusability principles, all datasets lack the use of a data model as well as of standards for the representation of data and metadata. Moreover, the absence of a clear data flow that describes the provenance of data makes it difficult to integrate data and perform a multi-country analysis. Positively, data are open and may be reused for statistical purposes without requiring authentication to relevant websites.

From a data quality perspective, the attention has been posed on the *inherent data quality* characteristics of ISO 25012. All datasets positively met the feature of currentness with information updated daily at local and national level. This is an important step forward that may be also applied for routinely datasets, as generally medical data are provided one or two years after the collection, making it difficult for scientists to produce innovative and non-obsolete analyses. On the contrary, datasets lack of understandability as no detailed information are reported in terms of indicator definition and formula adopted to compute it. Moreover, the lack of data flow describing its collection, elaboration, aggregation and diffusion makes datasets hard to be accurate and traceable. This is also underlined in previous work [8,9] considering regional based systems where the lack of standardized criteria for data collection might influence the count of cases and tests performed. Finally, the majority of countries provide data distributed by territorial, gender and age ranges level. However, a non-homogeneous distribution is present across both indicators and countries analysed. This data quality feature is critically important for the purpose of the datasets as a coherent distribution may allow a cross-country analysis of the COVID-19 diffusion in Europe.

References

[1] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, *et al*. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2015;3:1-9.

[2] Inau ET, Sack J, Waltemath D, Zeleke AA. Initiatives, Concepts, and Implementation Practices of FAIR (Findable, Accessible, Interoperable, and Reusable) Data Principles in Health Data Stewardship Practice: Protocol for a Scoping Review. JMIR research protocols. 2021;10(2):e22505.

[3] Smerek MM. Assessing Data Quality for Healthcare Systems Data Used in Clinical Research. 2015. Available at: https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/Assessing-data-quality_V1%200.pdf. Accessed July 28th, 2021.

[4] WHO. Global surveillance for COVID-19 caused by human infection with COVID-19 virus: Interim guidance. Available at: https://apps.who.int/iris/rest/bitstreams/1272502/retrieve. Accessed 14 July 2020. Accessed July 28th, 2021.

[5] ISO/IEC 25012:2008 - Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model, International Organization for Standardization, Switzerland. 2008.

[6] Chen Q, Allot A, Lu Z. LitCovid: an open database of COVID-19 literature. Nucleic acids research. 2021;49:D1534-D1540.

[7] Chen, H, Hailey D, Wang N, Yu P. A review of data quality assessment methods for public health information systems. IJERPH. 2014;11(5):5170-5207.

[8] Sartor G, Del Riccio M, Dal Poz I, Bonanni P, Bonaccorsi G. COVID-19 in Italy: Considerations on official data. Int J Infect Dis. 2020;98:188-190.

[9] Alamo T, Reina DG, Mammarella M, Abella A. Covid-19: Open-data resources for monitoring, modeling, and forecasting the epidemic. Electronics. 2020;9(5):827.