Producing Generative Digital Data Objects: An Empirical Study on COVID-19 Data Flows in Online Communities

Caroline Blotenberg
Freie Universität Berlin
c.blotenberg@fu-berlin.de

Arthur Kari
Freie Universität Berlin
a.kari@fu-berlin.de

Björn Kral
Freie Universität Berlin
b.kral@fu-berlin.de

Philipp Nürnberger
Freie Universität Berlin
p.nuernberger@fu-berlin.de

Hannes Rothe
ICN Business School
hannes.rothe@icn-artem.com

Abstract

Digital data objects on viruses have played a pivotal role in the fight against COVID-19, leading to healthcare innovation such as new diagnostics, vaccines, and societal intervention strategies. To effectively achieve this, scientists access viral data from online communities (OCs). The social-interactionist view on generativity, however, has put little emphasis on data. We argue that generativity on data depends on the number of data instances, data timeliness, and completeness of data classes. We integrated and analyzed eight OCs containing SARS-CoV-2 nucleotide sequences to explore how community structures influence generativity, revealing considerable differences between OCs. By assessing provided data classes from user perspectives, we found that generativity was limited in two important ways: When required data classes were either insufficiently collected or not made available by OC providers. Our findings highlight that OC providers control generativity of data objects and provide guidance for scientists selecting OCs for their research.

1. Introduction

At the beginning of 2020, the World Health Organization (WHO) declared the COVID-19 outbreak an international public health emergency [1]. As of writing, the pervasive consequences of the pandemic are omnipresent. Scientists intend to improve their understanding of these impacts to respond appropriately and mitigate the transmission of the virus. Actors in these international research efforts include laboratories, hospitals, medical centers, pharmaceutical companies, universities, and research centers [2]. Some are established actors in the viral domain, whereas others have only emerged as a response to the pandemic [3]. These entities aim to combat the COVID-19 pandemic by transforming viral data to digital data objects that help them innovate, e.g., in finding new diagnostics, vaccines, and treatments for the novel virus [4]. To achieve their goals, they rely on viral data collected by laboratories that are unaware of these goals, but still share their data fit-for-purpose in online communities (OCs). Due to diverse interests between actors who collect, share, and utilize COVID-19 data, only few standards on data apply [5, 3]. Ultimately, this widely limits actors in their research efforts [6].

To comprehend how actors exchange data in OCs, how effective it is, and how it might be improved, we examine OCs that provide SARS-CoV-2 nucleotide data by illustrating their network structures and assessing their influence on shared digital data objects. We investigate whether the provided data allow for a fair representation of context and how the network structures within COVID-19 OCs influence this. Therefore, we pose the following research question:

How do scientific online communities shape digital data objects on SARS-CoV-2 and COVID-19?

Our findings shed light on the structures of OCs and the potential of viral data to produce innovation, which is reflected in the concept of generativity. We also provide useful insights for the current and future viral outbreaks. The paper initially reviews the social-interactionist view on generativity in OCs. Thereupon we describe our explorative methodological approach, which covers data collection, integration, in-depth analysis, and visualization. Afterward, we discuss and reflect on our findings on OC structures and digital data objects.

URI: https://hdl.handle.net/10125/80154
978-0-9981331-5-7

2. Related Work

OCs enable an exchange of knowledge between actors that could lead to new insights and innovations. These can positively or negatively influence generativity. In the following sections, we will address the concepts of OCs and generativity in greater depth.

2.1. Online Communities

From the perspective of information systems research, digital infrastructures are critical for innovation [7]. In contrast, the social-interactionist viewpoint emphasizes that technical artifacts are not responsible for generating knowledge flows, instead it is the exchange of various audiences on these infrastructures that creates value [8].

An OC is a subset of digital infrastructures [9]. According to Faraj [10, 11], OCs can be defined as virtual organizations where actors, which are not necessarily known to each other, share common interests and benefit both personally and as a cooperative. The OCs we consider can be defined as digital platforms where members interact with each other in order to generate new knowledge and innovate through dynamic knowledge flows. Collaboration is an essential component of OCs, as this enables recombination and synthesis of expertise among their participants [3]. For our particular research endeavor, it is important to emphasize that knowledge flows between actors are unilateral and defined by data sharing. As data usage is not comprehensively tracked by OCs, it is distinguishable from reciprocal relationships, e.g., social networks [12].

2.2. Generativity of Digital Data Objects

Digital data, so-called bitstrings [13], represent an instance of interest based on *a priori* defined data classes [14]. In this context, data classes are an assemblage of attributes and relationships that are specified to digitally represent an instance object. Data instances are digital data objects and derive from their respective data classes. Digital innovation is highly dependent on encoding physical objects into a digital format and thus rendering them ”programmable, addressable, sensible, communicable, memorable, traceable, and associable” [15]. As a result, digital data is separated from the physical object [15] and is generally believed to be portable and reusable among many contexts [16, 17]. These characteristics speak to the notion of data being generative. Generativity is defined as ”a technology’s overall capacity to produce unprompted change driven by large, varied, and uncoordinated audiences” [18, p. 1980]. It describes the ability to generate innovation or new output beyond the intended purpose of the actor’s interactions [19, 20, 21]. Consequently, digital data is generative when it allows reuse beyond its initial purpose. Generative data therefore needs to represent data instances with a variety of attributes that users of such data can pick from to create new applications.

The social-interactionist view portrays generativity as a process through which knowledge flows as actors interact in an OC. Faraj [10] observed that interactions in an OC could either have a positive or negative impact on creating new knowledge. When dialogue stalls, e.g., when individual actors withhold knowledge or exit the OCs’ network, this has a limiting effect on the interactions [22, 23]. Whereas in generative exchanges, new knowledge outcomes are promoted [24]. In an OC in which actors access and share data, generativity materializes in the digital data objects. Actors argue and form consensus on what data instances are considered useful for future purposes and what data classes are relevant for these, yet unknown, endeavors.

Healthcare provides a societally important context in which sharing data is pivotal. Research has produced key innovations at outstanding speed to fight COVID-19 since the beginning of the outbreak [1]. Within a few months, scientists generated a vast amount of global COVID-19 bio data that flowed into OCs. Nevertheless, if data classes and instances are not appropriately managed, shared, and analyzed, these achievements are not fully utilized [2]. Generally, scientists collect data instances relevant to their specific research. These might appear incomplete for other scientists with deviating research endeavors, preventing the data from being widely used. However, if they collect and share complete data instances, it enables their peers to reuse the data beyond its initial purpose. This implies that data completeness promotes generativity.

The data instances contain collected bio data and their respective metadata. Bio data, such as nucleotide sequence strings of a specific virus [25] are described by data classes, oftentimes referred to as metadata. Metadata adds structured, rich contextual information to the respective bio data, and is thereby promoting reuse, aggregation, and integration of disparate data sets [5]. In this context, it includes, e.g. location (*where* was the sequence collected), time (*when* was the sequence collected), sequencing technology (*which* sequencing method was used), and host information (*from whom* was the sequence collected, e.g., homo sapiens), which are essential to unambiguously ascertain the context of the collection process. This allows researchers to compare analyses, evaluate outbreak progression and

variations in host specification [6].

To effectively ensure the completeness of data classes, life scientists call for a fit-for-purpose community standard for COVID-19 contextual data [26]. Improved interoperability between datasets and systems will enable improved consistency and ultimately empower new insights and discoveries in research [25]. In the literature, a few approaches already exist to provide consistent baseline requirements for published data. For instance, PHA4GE is a global coalition working to establish consensus standards for COVID-19 contextual data [5] and FORCE11 provides guiding principles for publishing accessible, interoperable, and reusable data [27]. An established metadata standard can support OC providers and actors to ensure that interactions and resulting digital data are generative. Such standards may have guiding effects but are neither enforced by research institutions, funding bodies, nor OC providers.

Due to technological advances, the amount of bio data is increasing exponentially [28]. Despite being mostly collected in research, data are frequently used by private companies [29] - e.g., for vaccine research. The mere number of provided nucleotide sequences is insufficient for considering the generativity produced within an OC, as the context is critical to ensure informative theory-building [28]. The completeness of data classes has an impact on data generativity, thereby influencing innovations that emerge from interactions in OCs [27]. According to our interviews, timeliness of collected and uploaded instances can promote its scientific use and thus also the generativity within the OCs. Therefore, we assume that the number of data instances, data timeliness, and data completeness can promote generativity, and thus are antecedents of generativity. In the following, we measure and evaluate these aspects.

3. Methodology

In this study, we collected metadata of nucleotide sequences (data instances) across eight different OCs to evaluate their impact on the antecedents of generativity and to explore the exchange of data among OCs. To accomplish this, we selected OCs based on current literature, in particular Bernasconi et al. [3]. Datasets from these OCs were downloaded and integrated, due to different schemas (data classes), using a Python-based data pipeline. This enabled a structured and comparable weekly integration process. Additionally, we performed a backward search within our dataset to ensure that all OCs were included. The resulting integrated dataset served as a single source of truth for the subsequent analysis. To substantiate claims and support our sensemaking from data, we discussed preliminary findings in 45- to 60 minute long interviews with life scientists, i.e., a leader of a German science laboratory working with COVID-19 data, the manager of a COVID-19 data platform in the UK, and a scientist from ELIXIR - a non-governmental research infrastructure provider.

3.1. Data Sources

In this section, we briefly introduce the examined databases. The selection consisted of open databases, providing at least 1,000 SARS-CoV-2 data instances. In order to meet our definition of an OC, databases had to allow actors to interact with each other. Thus, we only chose those offering a download and upload function.

GenBank NCBI is an extensive international DNA database and is part of the International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org/). This association includes other databases such as DDBJ, EMBL-EBI, and NCBI. For our analysis, the data was downloaded exclusively from GenBank NCBI (https://www.ncbi.nlm.nih.gov/sars-cov-2/), including data from EMBL-EBI and DDBJ, as they synchronize their data daily [30].

GISAID has the mission to enable a fast exchange of data on specific influenza viruses and SARS-CoV-2 (GISAID, https://www.gisaid.org/). The data is open and accessible after creating an account on GISAID.

CNCB provides access to relevant data for the life and health sciences (CNCB, https://bigd.big.ac.cn/). This database has a more national focus and provides SARS-CoV-2 sequences mainly, but not exclusively, from China.

VIPR specializes in viral data. Currently, it contains data of 20 different virus families, such as SARS-CoV-2 (VIPR, www.viprbrc.org). It covers a variety of countries and integrates data from other sources, such as Genbank NCBI.

COG-UK was established in the context of the COVID-19 pandemic and is operated by NHS organisations and the Wellcome-Sanger Institute. It contains UK-derived genomes of SARS-CoV-2 (COG-UK, https://www.cogconsortium.uk/). The OC provider itself as well as 12 other academic institutions take part in the sequencing [31].

CoV-Seq is a Chinese database that aggregates nucleotide data from multiple sources. Similar to COG-UK, CoV-Seq was initiated as a response to the COVID-19 pandemic. Its goal is to quickly provide SARS-CoV-2 genomic data for public use. It is regularly updated and cleaned by the integrator to enable researchers with little programming skills to use the data for their research (CoV-Seq, http://covseq.baidu.com/browse).

VirusDIP serves as an integrator providing viral data from CNGB, GISAID, and NCBI (VirusDIP, https://db.cngb.org/virusdip/ncov). The data originates from all over the world.

NMDC includes data from different microorganisms and seeks to assist the scientific community in making microbiological data more accessible for research (NMDC, https://nmdc.cn/nCov/en).

3.2. Integration of Data Sources

While there are several proposals for a uniform, global metadata standard, none are widely accepted and implemented [6]. We discovered that OCs establish their own data class standards by predetermining which data class attributes must be contributed and can be accessed publicly. It was observable by different structures regarding the number of given attributes, data specifications, and formats.

We gathered data instances weekly between 11/13/2020 and 03/02/2021, where data records were provided as JSON, CSV, TSV, or XLSX. Wherever possible, data instances were retrieved via provided APIs, although we had to automate data scraping with Selenium framework for GISAID, GenBank NCBI, and VIPR to get the respective files. To harmonize the different data formats, we integrated them into a MongoDB instance. Subsequently, we removed attributes that contained only one characteristic, or redundant information. Furthermore, we dropped entries that did not contain any meaningful information, such as the expression "?". The final output consisted of 36 data class attributes and approximately 1.6 million data instances in the first week, which grew steadily over time. Since some OC providers modified their data class by removing, modifying, and adding attributes, we updated the code correspondingly.

3.3. Table for Evaluation of Data Classes

To assess and compare how each OC contributed to data generativity, we created a table that specifies the percentage of entries containing meaningful information for each particular attribute. All attributes were assigned to one of four categories: sequence context, collection context, host context, and publication context. Furthermore, we classified the attributes according to their relevance for researchers in three types: optional, recommended, and required. The classification adheres to the PHA4GE specifications (www.pha4ge.org), which is a "COVID-19 Metadata Template" [32]. We set the attributes that were not covered in the template to optional.

3.4. Visual Network for Analysis of Community Structures

To investigate the data flow among actors, we conducted an exploratory network analysis using NetworkX (https://networkx.org/). We utilized attributes with information about laboratories that collected and submitted sequences to the OCs. The fields contained information such as the name and address of the respective laboratory. Since this information was obtained from free-form text fields, they varied in their level of detail, contained typos, and different abbreviations for the same laboratory.

To address these ambiguities, we removed leading and trailing spaces. After this, we transformed all strings to lowercase, converted special characters into an equal encoding format, and then divided the strings into substrings according to specific information content. To condense alternative designations, we extracted 1,300 lab name labels and manually mapped them to a unique ID if they referred to the same institution. This mapping constituted the basis for a training and testing set. 70% of the data formed the training set and 30% the testing set. This dataset was enriched with many non-matching lab names and randomly generated typos, resulting in a test dataset of 258,659 lab name combinations in total. To handle the constant addition of new lab name labels over time, we trained a Sklearn random forest classifier. It matches equal substrings of labs based on the Levenshtein Distances. On the test set, the classifier achieved an accuracy of 99.8%, recall of 83%, true-negative of 256,643 lab name combinations, false-positive of 14, false-negative of 284, and true-positive of 1,718. Due to the random forest classifier's recall of 83%, sporadic classification errors are inevitable.

Each node in the network either represents an originating lab, a submitting lab, a hybrid lab (labs that collect and submit sequences), or an OC. Laboratories that are labeled as "unknown" because of missing or incorrect information are represented in a separate node.

The edges represent a connection between nodes that exchange information with each other. For each node, we calculated the in-degree and out-degree centrality and correlated them to the contributed completeness of data classes. The final network in Figure 1 maps the data flow between actors.

4. Results

In this part, we outline the results of our research. To understand which actors collect and upload data instances, we visualized the data flows between them in a directed network. We discovered that the interactions of the actors determine the data generativity in the respective OC.

4.1. Community Structures

Figure 1 depicts the entire directed network of 2,303 actors of the eight OCs and an exemplary representation of a specific data flow. These actors are further classified as originating labs, submitting labs, and hybrid labs. There are 202 submitting labs, 1,601 originating labs, and 500 hybrid labs in the network. They are depicted as nodes and differ in color. The edges represent the flow of data instances between actors and adopt the color of its origin node. Additionally, we optimized the lengths of the edges using Fruchterman-Reingold's force-directed algorithm so that actors exchanging a large number of data instances are positioned closer to each other [33].

The network consists of significantly fewer submitting labs than originating labs. Submitting labs connect the originating labs to the OCs. The graph shows submitting labs near the center that sprout numerous edges. They combine data instances from multiple originating labs and transmit a large number of data instances to the OCs. Some actors connect to multiple OCs, indicating that they either upload to multiple OCs or synchronize their data with other OCs. Redundancies emerge in both cases. For four OCs, no structure was discernible since they did not specify originating or submitting labs.

In an evolving pandemic, data timeliness is important because interventions need to be implemented and evaluated quickly. Hence, data must be shared as fast as possible with anyone who may be able to use it [34]. Therefore, we assume that the timeliness of data instances influences generativity positively. The average time it takes to upload a collected data instance to a corresponding OC is approximately 73 days, although significant differences exist between individual labs and individual OCs. Though there is no significant correlation between the timeliness and completeness of the data, the labs' assessment in these dimensions allows selecting labs that perform well.

OC providers control data classes by designing database schemas for lab interactions. However, it is the responsibility of the individual labs to collect and transfer the instances according to these classes. We found that individual actors upload data instances that do not populate all the class attributes. The majority of labs that contributed a large number of data instances had filled required data class attributes between 60% and 70% on average. Additionally, they were faster than the arithmetic mean with a transmission time between 40 and 70 days. Lighthouse Labs particularly stood out in this regard. These are high-performance COVID-19 labs funded by the UK Department of Health and Social Care (DHSC). Their performance can be presumably attributed to their 24-hour operation, and highly automated processes [35][36].

4.2. Data Generativity in Online Communities

The interactions of all known and unknown actors in an OC determine the quantity and completeness of data instances. Hence, we extend our consideration and investigate the antecedents of generativity within the selected OCs. As generativity increases by sharing data within communities, it decreases with the cutback in available data instances [11]. Actors sequenced and published a significant portion of all positive COVID-19 cases at the start of the pandemic. However, the rate has decreased ever since. On 03/18/2020, actors sequenced 40% of reported COVID-19 cases, whereas it steadily declined to 1.5% since 01/05/2021. Consequently, the variation of data instances decreases, which limits generativity of data within OCs. In addition to decreasing sequencing rates, the OCs differ in their regional focus of coverage. For example, COG-UK only provides data instances that actors collected within the UK, while GISAID covers a variety of different countries. As a result, comparability for scientists using only a few OCs is limited. Hence, scientists attempt to combine data from a large selection of OCs.

To comprehend the exchange between OCs, we analyzed the data flows and determined that GISAID is a primary data source for several other OCs. This is an indicator for the combination of data from different OCs to increase the number of data instances and variation of data classes. This promotes generativity [21]. As a consequence and through the provision of APIs, the "generative engagement with others on a global scale" [21] is enhanced [11]. While it can have a positive impact, it also causes redundancy from duplicate data instances. The lack of a standard unique identifier prevents data records from being distinguished from one

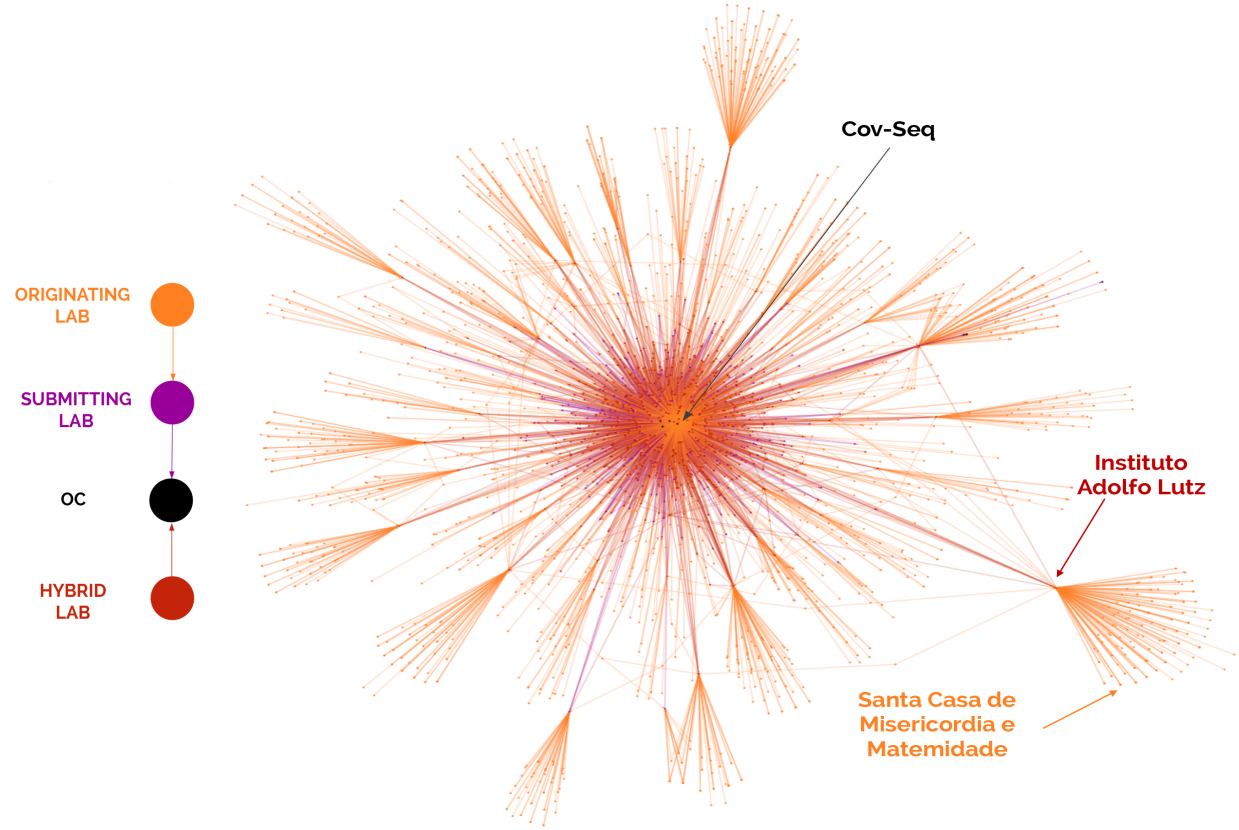

Figure 1. Network visualization of community structures

another and impedes data aggregation. Furthermore, the data completeness can be affected, as attributes might get lost during the exchange due to deviating data classes. For instance, the integrator ViruSurf provides a more extensive range of attributes of COG-UK data than COG-UK itself. This compromises generativity, as the completeness of data is significantly reduced by attributes that are not made publicly available.

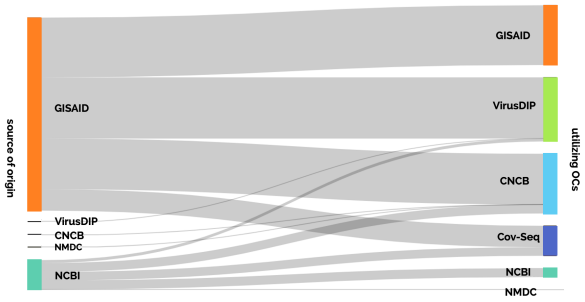

Figure 2. Data flows from source OC to utilizing OCs

In Figure 2, the cumulative data flows between OCs are shown. While APIs enable sharing of data among OCs, additional new data instances are generated every week by actors of the OC. For COG-UK, we observed a sharp increase in submitted data at the beginning of February 2021 that is presumably related to the emergence of mutations in the UK and the resulting intensification of sequencing. During our research, CoV-Seq was the only OC not contributing any new data instances. Furthermore, OCs update their data at different intervals: For example, GISAID and VIPR updated data daily whereas Genbank NCBI provides new data only once a week. In our observation period, the datasets in the respective OCs increased as shown in Figure 3.

However, we also noticed a removal of already submitted instances. Over three weeks, various OCs deleted a particularly large number of data instances at least once. CNCB and GISAID consistently removed several hundred records during these weeks. To explore the reason for the deletions, we performed a Chi-Square test for attribute selection utilizing the Sklearn library. In GISAID, we identified the attributes ”originating lab” and ”authors” as the most significant attributes for deleted entries whereas in CNCB they were ”last update

Table 1. Table for evaluation of data classes and their respective completeness

		Online Communities								Integrator
Relevance	Attribute	CNCB	Cog-UK	Cov-Seq	Genbank	GISAID	NMDC	VIPR	VirusDIP	ViruSurf
	Nucleotide Sequences	315,000	164,104	158,125	47,434	315,253	43,042	41,612	338,113	194,703
Sequence Context										
required	Strain Name	100.0%	100.0%	85.7%	-	100.0%	99.9%	100.0%	100.0%	99.0%
optional	Accession ID	100.0%	-	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
(optional)	Nextstrain Clade	-	-	-	-	100.0%	-	-	-	-
(optional)	GISAID Clade	-	-	-	-	100.0%	-	-	-	-
(optional)	Lineage	99.5%	88.7%	-	-	-	-	-	-	99.6%
(optional)	Lineage Support	-	88.7%	-	-	-	-	-	-	-
(optional)	Mol Type	-	-	-	-	-	-	100.0%	-	-
optional	Length	100.0%	-	100.0%	100.0%	100.0%	99.4%	100.0%	100.0%	100.0%
(optional)	Is complete	-	-	-	100.0%	-	-	100.0%	-	79.7%
(optional)	GC%	-	-	-	-	-	-	-	-	100.0%
(optional)	N%	-	-	-	-	-	-	-	-	100.0%
Collection Context										
required	Collection Date	98.9%	100.0%	96.6%	90.8%	99.0%	90.8%	97.5%	98.6%	98.7%
required	Location	100.0%	100.0%	97.5%	91.6%	100.0%	90.6%	-	99.6%	100.0%
required	Country	-	100.0%	97.5%	-	100.0%	-	100.0%	-	100.0%
recommended	Detailed Exposure Location	-	-	-	-	100.0%	-	-	-	-
recommended	Exposure Country	-	-	-	-	100.0%	-	-	-	-
required	Originating Lab	90.8%	-	-	-	100.0%	-	-	95.5%	60.0%
optional	Address Originating Lab	-	-	-	-	-	-	0.0%	-	-
optional	Email Originating Lab	-	-	-	-	-	-	0.0%	-	-
recommended	Assembly Technology	-	-	-	-	-	-	-	50.2%	20.3%
(optional)	Sequencing Technology	-	-	-	-	-	-	-	99.4%	20.7%
(optional)	Isolation Source	-	-	-	22.3%	-	42.6%	-	-	20.5%
(optional)	Coverage	-	-	-	-	-	-	-	-	1.7%
Host Context										
required	Host Information	100.0%	-	-	91.8%	100.0%	90.8%	98.5%	99.6%	100.0%
optional	Host Taxon ID	-	-	-	-	-	-	-	-	100.0%
required	Host Taxon Name	100.0%	-	-	91.8%	100.0%	-	98.5%	-	100.0%
recommended	Host Age	-	-	-	-	21.9%	-	-	-	6.6%
recommended	Host Gender	-	-	-	-	23.4%	-	-	-	6.5%
Publication Context										
required	Submitting Lab	94.9%	-	85.7%	-	100.0%	-	-	100.0%	81.7%
optional	Email Submitting Lab	-	-	-	-	-	-	0.0%	-	-
optional	Address Submitting Lab	-	-	-	-	-	-	0.0%	-	-
recommended	Submitting Author	-	-	99.3%	99.1%	-	0.4%	-	99.9%	-
optional	Title of Publication	-	-	-	100.0%	8.1%	0.0%	-	1.2%	-
optional	Submission Date	100.0%	-	-	100.0%	100.0%	100.0%	-	100.0%	81.8%
(optional)	Biosample ID	-	-	-	53.5%	-	-	-	-	11.0%
(optional)	Bioproject ID	-	-	-	-	-	-	-	-	11.0%

time" and "create time". In GISAID, the deletions of several hundred data instances can be retraced to two laboratories. In CNCB, several hundred data instances that had all been transmitted from a specific lab at the exact same minute were deleted.

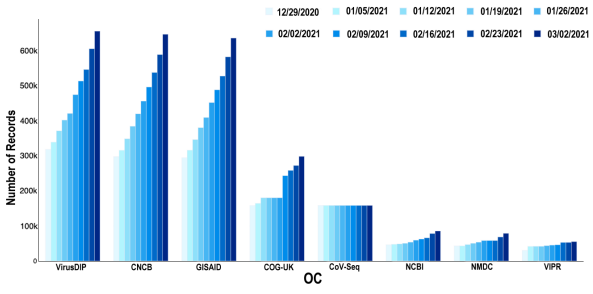

Figure 3. Number of records over time

We assume that errors in the collection or transmission process are responsible for these deletions. Said removals harm generativity as they prevent the reproducibility of research results for scientists and reduce the amount of publicly available data instances. However, the positive effects of deleting erroneous data can potentially outweigh the loss of generativity if it prevents future research from being faulty. Besides, the OC providers violate the integrity of OCs by significantly changing data classes. We observed that COG-UK added new attributes identifying evolving mutations. Thus, OC providers respond to developments in the pandemic and give additional information that contributes to generativity by making it easier to find and utilize relevant data instances. However, this can lead to errors in automated pipelines.

We examined 36 attributes of nine data classes grouped by sequence context, collection context, host

context, and publication context for each OC in terms of their completeness. The displayed importance of each attribute reflects its relevance for research purposes. OCs differ vastly in the combination of attributes included in their classes, as Table 1 illustrates. For instance, GISAID offers 18 attributes while covering most of the required information. On the contrary, COG-UK only contains six attributes while missing required information about the host, originating lab, and submitting lab completely. A comparison between the integrator ViruSurf and the OCs shows that various attributes seem to be transferred to integrators, but are not made publicly available.

It is also evident that a large quantity of data class attributes is either not given at all or only given in a few data instances. This hinders comparability between data instances, deems the classes unusable, and therefore inhibits generativity of data in the OC. Generally, the ”required” class attributes are more frequently filled with meaningful information compared to ”recommended” and ”optional” attributes. Although they are assumed to have a decisive influence on innovations within OCs [27], they are completely missing in numerous OCs.

Our complementary interviews revealed that scientists rely on a combination of different relevant data class attributes depending on the research question they address. Consequently, the completeness of individual attributes is not sufficient to consider how many data instances can be used for research. When combining a selection of required attributes, a progressive flattening of remaining data instances can be observed in Figure 4. It demonstrates that data loss is very high, with only 12% of the original data instances persisting. Furthermore, if the remaining

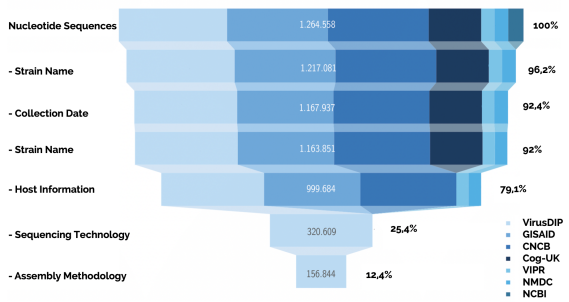

Figure 4. Loss of data instances when filtering for (relevant) attributes

required attributes were added to the funnel, there would be no data instances left. Thus, depending on how many attributes a researcher requires, only a minor part of the original data instances would be usable. Nonetheless, if ”sequencing technology” and ”assembly method” were omitted, approximately 79% of data remain. The completeness of the individual data class attributes changes over time. At the beginning of the pandemic, the attribute ”host information” was specified in 95% of all data instances. With the increase in global sequencing in April 2020, the proportion of records containing host information dropped to approximately 73% and now remains at almost 75%. We demonstrated that each OC represents a network of actors and their interactions with one another. Each OC provider establishes a discrete comprehension of data classes and thus influences data generativity that results from the participants' interactions. They also impact generativity by making adjustments to their data classes, impeding comparability of research results. However, actors also limit generativity by reducing sequencing rates, delaying transmission, and not adhering to data classes by leaving out significant attributes that promote the reusability of data.

5. Discussion

This paper speaks to the social-interactionist view on generativity [11]. Its empirical foundation was mainly limited to products and platforms [37]. According to Faraj [11], the interactions of the participants are the underlying antecedents of generativity. OCs have fluid boundaries and a dynamic virtual space in contrast to typical organizational hierarchies [10]. In our case, OC providers control data classes and thereby shape interactions of their participants. Instead of dynamic data flows, we observed an exchange of pre-defined data instances. By studying data flows within OCs, we extend the social-interactionist view on generativity to digital data objects. We claim that generativity materializes in the digitized sequence itself and is extended if the digital data object is enriched with additional contextual information of the physical sequence on the OC. This is important, because the more contextual information is associated with a data object, the more it might meet the needs of future innovators who seek to innovate with said data objects using yet unknown means in yet unknown subject areas. The digital data object can be shared more rapidly with actors around the world than the physical object it originates from. In light of the ongoing pandemic, if more researchers can share and access large quantities of those digital instances quickly, more actors can simultaneously work on life-saving solutions against novel viruses.

Our study provides insights on a mismatch between

how data is collected and how it is shared, which influences generativity. Data is created for particular purposes, which have an impact on the means that produce data and how it is organized. We found that sequences from labs were frequently derived from research projects examining specific research questions. As a result, only information pertinent to research objectives were retrieved from the sequences. However, OC providers define static data classes through class-based interface designs that contain only a small number of attributes. Consequently, if collected instances do not match the class-based interfaces, labs might abdicate data entry efforts. Thus, additionally collected information is lost. However, an instance-based approach would lead to the loss of the consumer orientation since collected data might fluctuate vastly in content and level of detail [14]. Consumers of the resulting digital data might be limited in their ability to aggregate and analyze large amounts of data. As a result, the generative properties of digital data might be attenuated from this reorganization. Our data thereby suggests that instance-based collection of nucleotide sequences and class-based sharing of these within OCs can significantly limit generativity of digital data. To maximize the fit-for-purpose and ensure heterogeneity of shared digital data, a complementary contributor-centric approach [14] could be implemented that selectively opens categories to unexpected data. One example are NMDC queries with the attribute "host" in contrast to narrow class attributes such as "host taxon name".

This study also carries implications for health policy and management. Data accessibility is critical for research, particularly when dealing with an existential threat to humanity, such as SARS-CoV-2. Our results revealed that data availability differs significantly among OCs as not all countries are represented. Thus, the global coverage of SARS-CoV-2 is limited. Data availability is also affected by a lower rate of sequencing since labs at the time of writing only sequence 1% of positive cases. To increase this rate and hence generativity, monetary incentives can be provided. For instance, the Federal Ministry of Health in Germany defined a 5% target line in their SARS-Cov-2 surveillance order, ensuring a payment of 220 euro per sequence to laboratories [38]. Since the virus spreads fast and mutations occur, new data instances must be uploaded as soon as possible [39]. However, we discovered that it takes an average of 73 days for collected samples to be uploaded. To accelerate this process, decision-makers could link the recency of data to monetary incentives so that labs have inducements to establish higher capacities [40]. Over the course of several weeks, we observed deletions of data instances, changes in data classes, and OC providers limiting the number of publicly available data class attributes. In addition, three of the selected OCs did not provide an API. These issues can lead to limitations in downloading and fully utilizing the digital data object, which harms generativity. Instead of deleting data, OC providers could add attributes displaying potentially flawed data records. To avoid ambiguities in e.g. lab names, OC providers should implement an auto-suggest feature that proposes already existing lab name spellings to contributors.

This research is subject to limitations. Since the data used in our research does not contain information about actors that download data, the results do not reveal how scientists use data instances. While being able to assess how OCs shape digital data objects and thus influence the antecedents of generativity, we were unable to evaluate the actual effects on COVID-19 research. Furthermore, substantial redundancies occur because OCs exchange data with each other. Thus, the analysis includes duplicated data. Another limitation is the inevitability of classification errors of the presented random forest classifier, that aims to match ambiguous lab names. These errors can result in sporadic incorrect assignments of data entries to individual labs.

References

[1] European Commission, "Horizon 2020 projects working on the 2019 coronavirus disease (COVID-19), the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and related topics: Guidelines for open access to publications, data and other research outputs," *Research and Innovation*, vol. 2, 2020.

[2] A. Smith, C. Martin, N. Blomberg, and ELIXIR Consortium, "ELIXIR Position Paper on Suitable Business Models for Open Data [version 1]," *F1000Research*, vol. 9, no. 65, 2020.

[3] A. Bernasconi, A. Canakoglu, M. Masseroli, and P. Pinoli, "A Review on Viral Data Sources and Integration Methods for COVID-19 Mitigation," Aug 2020.

[4] European Commission, "EU research and innovation in action against the coronavirus: funding, results and impact.," *Research and Innovation*, vol. 2, 2020.

[5] A. J. Emma Griffiths and R. E. Timme, "The PHA4GE SARS-CoV-2 contextual data specification for open genomic epidemiology behalf of the Public Health Alliance for Genomic Epidemiology (PHA4GE) consortium," Aug 2020.

[6] L. M. Schriml, M. Chuvochina, N. Davies, E. A. Eloe-Fadrosh, R. D. Finn, P. Hugenholtz, C. I. Hunter, B. L. Hurwitz, N. C. Kyrpides, F. Meyer, I. K. Mizrachi, S. A. Sansone, G. Sutton, S. Tighe, and R. Walls, "COVID-19 pandemic reveals the peril of ignoring metadata standards," Dec 2020.

[7] M. Shaikh and E. Vaast, "Folding and unfolding: Balancing openness and transparency in open source

communities," *Information Systems Research*, vol. 27, pp. 813–833, Sep 2016.

[8] S. Jarvenpaa and W. Standaert, "Digital Probes as Opening Possibilities of Generativity," *Journal of the Association for Information Systems*, vol. 19, no. 10, p. 3, 2018.

[9] P. Spagnoletti, A. Resca, and G. Lee, "A design theory for digital platforms supporting online communities: A multiple case study," *Journal of Information Technology*, vol. 30, no. 4, pp. 364–380, 2015.

[10] S. Faraj, S. L. Jarvenpaa, and A. Majchrzak, "Knowledge collaboration in online communities," *Organization Science*, vol. 22, pp. 1224–1239, Sep 2011.

[11] S. Faraj, G. von Krogh, E. Monteiro, and K. R. Lakhani, "Online community as space for knowledge flows," *Information Systems Research*, vol. 27, pp. 668–684, Dec 2016.

[12] O. Hanseth and K. Lyytinen, "Design Theory for Dynamic Complexity in Information Infrastructures: The Case of Building Internet Organisational learning View project Communication and Design View project," *Article in Journal of Information Technology*, 2010.

[13] J. Runde and P. Faulkner, "Theorizing the digital object," *MIS Quarterly*, vol. 43, 08 2019.

[14] R. Lukyanenko, J. Parsons, Y. Wiersma, and M. Maddah, "Expecting the unexpected: Effects of data collection design choices on the quality of crowdsourced user-generated content," *MIS Quarterly*, 09 2018.

[15] Y. Yoo, O. Henfridsson, and K. Lyytinen, "Research commentary—the new organizing logic of digital innovation: An agenda for information systems research," *Info. Sys. Research*, vol. 21, p. 724–735, Dec. 2010.

[16] C. Alaimo and J. Kallinikos, *Encoding the Everyday: The Infrastructural Apparatus of Social Data*. 11 2016.

[17] S. Nambisan, K. Lyytinen, and Y. Yoo, *Handbook of Digital Innovation*. Research Handbooks in Business and Management series, Edward Elgar Publishing Limited, 2020.

[18] J. L. Zittrain, "The generative internet," *Harvard Law Review*, vol. 119, no. 7, pp. 1974–2040, 2006.

[19] O. Henfridsson and B. Bygstad, "The generative mechanisms of digital infrastructure evolution," *MIS Quarterly: Management Information Systems*, vol. 37, pp. 907–931, Sep 2013.

[20] M. de Reuver, C. Sørensen, and R. Basole, "The Digital Platform: A Research Agenda," *Journal of Information Technology*, vol. 33, Apr 2017.

[21] J. Zittrain, "The Future of the Internet and How to Stop It," tech. rep., 2008.

[22] G. C. Kane, J. Johnson, and A. Majchrzak, "Emergent life cycle: The tension between knowledge change and knowledge retention in open online coproduction communities," *Management Science*, vol. 60, pp. 3026–3048, Dec 2014.

[23] P. Vassilakopoulou, E. Skorve, and M. Aanestad, "Enabling openness of valuable information resources: Curbing data subtractability and exclusion," *Information Systems Journal*, vol. 29, pp. 768–786, Jul 2019.

[24] M. R. Kamdar and M. Dumontier, "An Ebola virus-centered knowledge base," *Database The Journal of Biological Databases and Curation*, vol. 2015, no. March, pp. 1–11, 2015.

[25] J. R. Brister, Y. Bao, C. Kuiken, E. J. Lefkowitz, P. le Mercier, R. Leplae, R. Madupu, R. H. Scheuermann, S. Schobel, D. Seto, S. Shrivastava, P. Sterk, Q. Zeng, W. Klimke, and T. Tatusova, "Towards viral genome annotation standards, report from the 2010 NCBI annotation workshop," *Viruses*, vol. 2, no. 10, pp. 2258–2268, 2010.

[26] S. Canham and C. Ohmann, "A metadata schema for data objects in clinical research," *Trials*, vol. 17, no. 1, pp. 1–11, 2016.

[27] FORCE11, "Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version b1.0," p. 1, 2014.

[28] B. Pentland, J. Recker, and G. M. Wyner, "Bringing context inside process research with digital trace data Quantitative Data Analysis: A Companion for Accounting and Information Systems Research View project Positive Deviance View project," *Article in Journal of the Association for Information Systems*, 2020.

[29] H. Rothe, S. L. Jarvenpaa, and A. Penninger, "How Do Entrepreneurial Firms Appropriate Value in Bio Data Infrastructures: An Exploratory Qualitative Study," 2019.

[30] INSDC, "International nucleotide sequence database collaboration," Mar 2021.

[31] W. S. Institute, "COVID-19 Genomics UK (COG-UK) Consortium," 2021.

[32] PH4GE, "COVID-19 Metadata Template," 2021.

[33] T. Fruchterman and E. Reingold, "Graph Drawing by Force-Directed Placement," *Softw., Pract. Exper.*, vol. 21, pp. 1129–1164, 11 1991.

[34] S. P. J. M. Horbach, "Pandemic publishing: Medical journals strongly speed up their publication process for COVID-19," *Quantitative Science Studies*, vol. 1, pp. 1056–1067, Aug 2020.

[35] A. Richter, T. Plant, M. Kidd, A. Bosworth, M. Mayhew, O. Megram, F. Ashworth, L. Crawford, T. White, E. Moles-Garcia, J. Mirza, B. Percival, and A. McNally, "How to establish an academic SARS-CoV-2 testing laboratory," *Nature Microbiology*, vol. 5, Dec. 2020.

[36] Department of Health and Social Care, Gov-UK, "Guidance NHS test and trace: How we test your samples," Feb 2021.

[37] C. Cennamo and J. Santalo, "Generativity tension and value creation in platform ecosystems," *Organization Science*, vol. 30, 05 2019.

[38] Bundesministerium für Gesundheit, "Referentenentwurf des Bundesministeriums für Gesundheit - Verordnung zur molekulargenetischen Surveillance des Coronavirus SARS-CoV-2," 2021.

[39] V. Moorthy, A. Restrepo, M.-P. Preziosi, and S. Swaminathan, "Data sharing during the novel coronavirus public health emergency of international concern," Jan 2020.

[40] G. Abi Younes, C. Ayoubi, O. Ballester, G. Cristelli, G. de Rassenfosse, D. Foray, P. Gaulé, G. Pellegrino, M. van den Heuvel, E. Webster, and L. Zhou, "COVID-19: Insights from innovation economists," *Science and Public Policy*, Jul 2020.