Different Kinds of Data: Samples and the Relational Framework

Aline Potiron, Johannes Kepler University, Linz, Austria

0000-0003-3521-882X

Abstract:

This paper proposes an original definition of samples as a kind of data within the relational framework of data. The distinction between scientific objects (e.g., samples, data, models) often needs to be clarified in the philosophy of science to understand their role in the scientific inquiry. The relational framework places data at the forefront of knowledge construction. Their epistemic status depends on their evaluation as potential evidence in a research situation and their ability to circulate among researchers. While samples are significant in data-generating science, their role has been underexplored in the philosophy of data literature. I draw on a case study from data-centric microbiology, viz. amplicon sequencing, to introduce specifications of the relational framework. These specifications capture the distinctive epistemic role of samples, allowing the discussion of their significance in the inquiry process. I argue that samples are necessarily transformed to be considered as evidence, portable in the limits of a situation, and they act as world anchors for claims about a phenomenon. I compare these specifications with other data and evidence frameworks and suggest they are compatible. The paper concludes by considering the extension of these criteria in the context of biobanking. The specifications proposed here help analyze other life sciences cases and deepen our understanding of samples and their epistemological role in scientific research.

Keywords: Samples, Data, Relational Framework, Amplicon Sequencing

Acknowledgements:

Earlier versions of this work were presented at the ISHPSSB in 2021 and the EASPLS in 2022, where it received many valuable comments. I thank the reading group of the Philosophy Institute at JKU, who also gave me helpful comments and feedback. I thank Emanuele Ratti and Julian Reiss for their careful readings. I am grateful to Sabina Leonelli for the early exchanges on this work and follow-up comments. I am also thankful to Sophie Veigl for her constructive advice on the methodological part. I thank Charlie Pauvert and Michalis Christou for helping proofread the manuscript. Finally, I created Figure 1 with Biorender.com.

Statements & Declarations:

No specific funding was received to conduct this study.

The author has no competing interests to declare relevant to this article's content.

1. Introduction

In the philosophy of science, data are objects participating in constructing scientific knowledge. Defining data implies understanding this role and how it potentially differs from the role of other scientific objects, such as samples, models, or theories. In philosophical discussions, data are linked to the old dichotomy between observation of the world and what we know about it. In this context, data have the power to constrain what we can say about the world. They are empirical constraints and connect our ideas of the world (theories, knowledge, assertions) to the world.

The relational framework of data places data at the center of the construction of scientific knowledge. It spurs from practice analysis in which theory plays, at best, an attenuated role. The status of a product of the research activity as data depends on its evaluation to function as potential evidence by the actors of a given research situation. And it also depends on its capacity to be disseminated – to travel – among these actors (Leonelli 2016).

Samples are also prominent scientific objects in data-generating research activities. Still, their analysis has been neglected in the philosophy of data literature (Leonelli and Tempini 2020, viii and 17), which calls for more clarity on their role and place in scientific inquiry. Indeed, current works sometimes place sample collection as part of "the early stages of a data journey" (Halfmann 2020, 27) or as part of the data (Wylie 2020, 298). Other times, samples are explicitly distinguished from data: "[I]t remains to be seen whether the model described here applies to sample journeys as well as to data journeys" (Griesemer 2020, 162, footnote 17).

Therefore, I propose an original and timely definition of samples as a kind of data within the relational framework. The argument proceeds via a case study representative of microbiology and, more generally, biology: the technique of amplicon sequencing.

Section 2 presents the relational framework of data, its motivations, definitions, and advantages in contemporary life sciences. Section 3 has two parts. First, I describe a case from data-centric microbiology, amplicon sequencing (AS). Second, I subscribe to the relational framework of data and explore its fitness for this case study. I agree that its criteria are essential for analyzing and understanding the epistemic role of scientific objects. Still, these criteria need to be more precise to make sense of the distinctive role played by samples in the case used here. Section 4 specifies the characteristics of samples[1] compared with other kinds of data within the relational framework of data. These specifications make the discussion of the significance of samples' epistemic role in the inquiry process possible. Section 5 compares these specifications with data and evidence frameworks developed since the relational

[1] These precisions and the role of samples in the inquiry described here should be compared with the definition of statistical samples. In philosophy of statistics, "samples" are actual values or data obtained in a specific setting. They are part of a whole, called the "sample space" constituted by all possible values in this setting (see Romeijn 2022). This comparison is beyond the scope of this article. I restrain my conclusions to the life sciences.

framework. In Section 6, I conclude with a possible generalization of these criteria in the case of biobanking.

2. The Relational Framework of Data

2.1. Motivations for a New Framework of Data

The concept of data is related to the difference between observations of the world and the knowledge we form about it. There is tension between what data are and what data do in knowledge construction.

The traditional view of scientific inquiry focuses on theories. Knowledge construction is about how they are formed, confirmed, or chosen when conflicting theories are available. Theories are conceptualized as sets of sentences in a formally structured language, a theoretical language. Comparing these sentences with what happens in the world needs observational reports—data—of the phenomenon expressed in an observational language (Boyd & Bogen 2021) and relations of interpretations between these two languages (Peschard & van Fraassen 2018, 23). Data represent a part of a phenomenon, and using the relations of interpretations and the theory, scientists arrive at an explanation of this phenomenon or the prediction of another phenomenon (Hempel 1952, 36). Data are considered reliable in this conception to test the theory's accuracy. They have a fixed, context-independent, and objective representational content. This view is also called the syntactic view of theories and formed the core of early 20th-century logical empiricism.

By the 1960s, the semantic view of theories introduced a new focus on models rather than theories. Patrick Suppes advocates an explicit focus on models to understand scientific representation. Models are indispensable tools for linking data and theories, which are understood as sets of models. In a later development of his view, he constructed a hierarchy of models from data to theory (e.g., "data models") to make this relationship explicit.

At the end of the 1970s, the focus on models contributed to what is sometimes called the "practice turn" in philosophy and sociology of science. The common point is a specific emphasis on actual scientific practice. The main aim is to give a detailed, meticulous, and descriptively adequate account of all stages of actual scientific practice and reasoning (Soler, Zwart, Israel-Jost, et al. 2014, 12).

During this turn, the purpose of scientific representations changes from a description of the world to a means for intervening in that world. Moreover, the "format" of these representations plays a role in the epistemic content they convey (Soler, Zwart, Israel-Jost, et al. 2014, 9, 16, and 24). Thus, data are considered valuable representations. They are not just theory-testing devices; they count as scientific achievements and can change locations to obtain new articulations of the world (Latour 1999, 306-7). They can be justified outside of, and be more robust than, theories; they can be retained while the theory is not. They have a fixed representational content, but this content is more context-dependent.

Sabina Leonelli developed the relational framework of data as an alternative to the available concepts described above. According to Leonelli, these concepts are either theory-centered,

"representation"-centered, or both. This does not match contemporary scientific practices in biology.

The syntactic view of theories is too theory-centered, whereas contemporary biology is "data-centered" (Leonelli 2016). Many data are generated in such practices, and there is a strong emphasis on their production, storage, manipulation, and dissemination. In addition, data participate in discovery; they are valuable outcomes of the scientific process and are thus worthy of public, scientific, and philosophical attention. The relational framework emphasizes this central place of data. It considers the iterative process between different phases of inquiry and the various products of this inquiry (Leonelli 2019, 18-25).

The "relational" view of data is constructed in opposition to what Leonelli coined the "representational" view of data (Leonelli 2016, 74). Syntactic and semantic views of theories constitute the latter. In both, data are representations: they capture some of the mind-independent properties of a phenomenon. They also have fixed and context-independent representational content. Data are conceptualized primarily as representations during and after the practice turn. The representational content is more context-dependent than the syntactic view but is still fixed. The relational framework calls into question the stability of the data content – whatever this can be: information, a signal, etc. It is relational because it stresses the role of the inquirer(s) in determining what counts as data. Data are defined "by the evidential value ascribed to them within specific research situations" (Leonelli 2016, 5). I subscribe to this framework for the remainder of the paper and for my case study analysis.

2.2. Definitions

In the relational framework, data are "any product of research activities, ranging from artifacts such as photographs to symbols such as letters or numbers, that is collected, stored, and disseminated *in order to be used as evidence for knowledge claims*." (Leonelli 2016, 77, original emphasis). Any particular object produced by a scientific inquiry can be data. They can be "any product of research activities […] that (1) are treated as potential evidence for one or more claims about phenomena and (2) are formatted and handled in ways that enable its circulation among individuals or groups for the purpose of analysis" (Leonelli 2016, 77-78). One can summarize these two criteria as 1) being considered potential evidence for knowledge claims and 2) having the capacity to travel between different situations of scientific inquiry. Let me unpack this definition.

First, data are mobile. The notion of "traveling" ("dissemination" and "circulation" are also used; see quotations above) builds on ideas developed by Howlett and Morgan (2011). In that work, the interpretation of traveling is quite literal. Facts travel in space and time, between disciplines, across epistemic traditions, etc. (Morgan 2011). Leonelli defines "traveling" as "data journeys." That is, "the movement of scientific data from their production site to many other sites within or beyond the same field of research" (Leonelli 2016, 39). This notion is a helpful metaphor for highlighting key features. These suggest the need for infrastructure (e.g., databases) that aids dissemination, appropriate vehicles (e.g., file formats, particular software, etc.), and financial resources (Leonelli 2016, 39-41). Even though data have no intention or

agency, they cannot plan, book, etc. (Leonelli 2016, 42). The notion of travel is literal. Data journeys "range from very concrete shifts of materials from one individual to another […] to highly diffused and depersonalized dissemination" (Leonelli 2016, 40). It is the displacement between at least two things (e.g., two different times, locations, disciplines, etc.).

Second, what count as data and the representational value of data are situation-dependent. Situationism (borrowed from John Dewey) is, roughly speaking, a kind of contextualism. The idea is that the "research context may have flexible and dynamic boundaries" (Leonelli 2016, 183-184). It gathers only those elements (whether they are events, objects, concepts, social features, etc.) of the context that are relevant to the agent's current inquiry (Brown 2012; Dewey 1938). No a priori elements are or are not necessarily part of a situation. For example, a biologist may use sequencing DNA to identify the type of microorganisms in an environment. The context includes external conditions, such as the weather, and background knowledge, such as the theory of evolution. They relate differently to the context. Suppose that the weather is known to influence the DNA sequencing process, leading to potential bias. In that case, it becomes a relevant part of the context and, thus, part of the situation. It should be documented in the metadata. Evolutionary theory is relevant in a different way. It can be relevant in interpreting the DNA sequences relative to the aim of the investigation. For example, it can help understand the evolutionary history of the environment or evaluate the

results in the context of evolution[2]. However, not all contextual factors are always relevant, and the biologist must carefully consider which factors are part of the situation.

Meanwhile, Dewey links the concept of a situation tightly to the idea of inquiry (Brown 2012). The inquirer's perception of an unsettled situation sparks the inquiry. Based on this perception, she will develop research questions. The investigator's judgment at the end of the inquiry will settle the new situation.

2.3. Advantages of the Relational Framework

The relational framework presents two main advantages. First, it fits particularly well with contemporary scientific practices such as data-centric science and has been used to analyze how data play their role in various scientific inquiries (e.g., Leonelli and Tempini 2020; Currie 2021; Pietsch 2015; Lloyd et al. 2022). Second, it clearly distinguishes between data and models.

The same data can be used to support various knowledge claims in the context of big data and data-centric practices. "[T]he same set of data can act as evidence for a variety of knowledge claims, depending on how they are interpreted—a feature that I take to be central to understanding the epistemic power of data as research components" (Leonelli 2016, 79). This

[2] I thank an anonymous reviewer to help me clarify the distinctive roles of different contextual elements within the situation.

is not easy to account for in representational frameworks because what data represent is fixed. Yet, one characteristic of the relational framework is that data have broad representational power and can represent different things depending on the research situation (Leonelli 2019, 19).

Moreover, Suppes' model hierarchy is very well suited to numerical and statistical data (Leonelli 2019, 7). However, it is unsuitable when the data are more varied (images, photos, etc.). Exploratory research often encountered in biology uses this type of data. The relational framework is better suited to this type of research because it has a broader understanding of the types of objects that data can be. The idea is that any product of a given situation *can* be considered as data. However, whether this is realized depends on many factors, including the investigator(s), the situation in which she is, the aim of the inquiry, etc. This does not mean that anything is data. Nor does it equate to a specific scientific product that can serve as evidence for any biological phenomenon (see Leonelli 2016, chapter 3, especially note 23). The criteria must be fulfilled.

The view inherited from Suppes suffers from seeing everything as a model. Yet this is an uncomfortable position (Leonelli 2019, 5). Indeed, if "data" are one type of "model," what makes that type so particular that it can judge which theory is "truer" or at least more empirically adequate than another? Leonelli (2019) argues that the relational framework provides a more precise distinction than the representational view. In the relational framework, models are a way to organize data to represent a targeted phenomenon. A phenomenon can be

anything considered to be occurring in the world. It is what scientists judge interesting to study[3]. Modeling or ordering of data narrows the actual representational meaning of data. Here, modeling has nothing to do with a higher degree of abstraction. Thus, it differs from Suppes' conception of "models of data." Instead, it is a reduction, a focus of the initial broad representational possibilities of data.

3. Different Kinds of Data

I argue that distinguishing several kinds of data within the relational framework helps better understand the methodology and epistemology of amplicon sequencing (AS) than the notion of data alone. This distinction allows us to understand the specificity of the role played by different objects, specifically the role of "samples." Through the description of AS (Figure 1),

[3] I place my analysis within the relational framework of data, so in this manuscript, I make no ontological commitment as to what phenomena can be. I follow Leonelli's definition: A phenomenon is the target of the scientific inquiry. It is "the target of the claims for which data can be used as evidence" (Leonelli 2019, 3), and it can be more or less well-defined at the beginning of the inquiry. I believe the choice of characterizations for a phenomenon may not significantly impact my analysis. For instance, my choice does not hinder me from drawing conclusions regarding samples and data compatible with existing literature (see citations in section 5).

I map the scientific products using criteria 1) and 2) of the relational framework (Table 1). I also highlight the specificity of different kinds of data in their role in scientific inquiry.

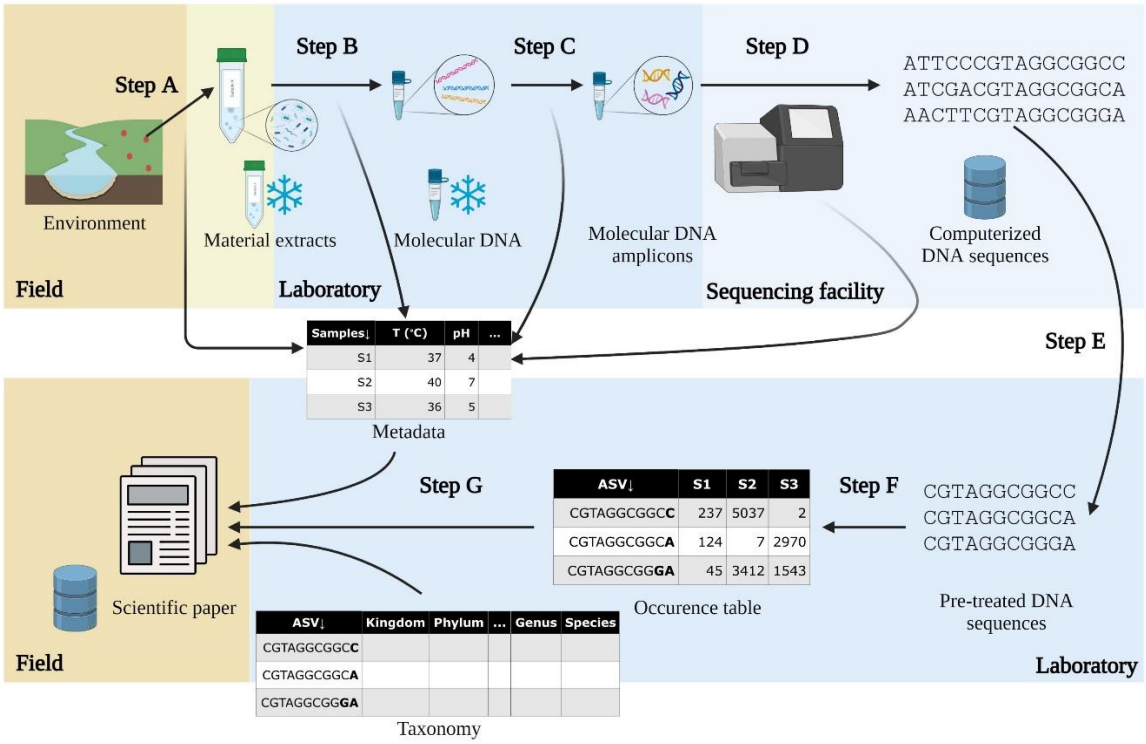

Fig. 1 Schematic representation of the seven steps of the amplicon sequencing method

3.1. Amplicon Sequencing

The AS method is widely used in microbiology and microbial ecology (Richardson et al. 2023). It is supposed to "reveal" the underlying characteristics of the microbial community in any environment (Quince et al. 2017). AS substitutes microbial observation by generating millions of DNA sequences. This data-centered method qualifies, in principle, for the relational framework of data.

My situated experience as a researcher[4] gave me knowledge of the steps of this procedure. It also helped me generate preliminary hypotheses about sample definition, data definition, and their epistemic roles in AS. Concepts in the philosophy of data challenged these hypotheses, and I investigated these hypotheses in a mixed-methods design—close reading of research papers and informal, semi-structured conversations with informants that were documented. I took these data as direct information and did not apply a particular analysis to them. I adopted something similar to an "iterative dialogue" (Mansnerus and Wagenknecht 2015, 45-46) between the concepts I was developing, informed by philosophical frameworks and the practice of the scientists.

On the basis of this analysis and published work (Alteio et al. 2021; Pollock et al. 2018), I decompose AS into seven steps, each yielding a scientific product. More specific examples come from the Earth Microbiome Project (EMP). The EMP is a collaborative project between researchers from around the world, and the aim is to understand microbial communities across the environment globally. To do so, data are gathered from individual studies (Thompson et al. 2017).

[4] I have performed a doctoral research project between 2014 and 2017 in the Micalis Institute at the INRAE of Jouy-en-Josas, France. This institute performs this analysis regularly. During this time, I performed myself steps A to C and was in direct contact with bioinformaticians for steps D to G It gave me a "feeling with" the phenomenon (Mansnerus & Wagenknecht 2015).

Step A – The environment of interest (e.g., soil) is sampled for analysis in a laboratory. These material extracts share properties with the world where they originate, e.g., the pH of the bulk of soil is similar to that of the environment where it originates. They are stored at least long enough to reach the laboratory. The context and specific means used to obtain the material extracts are stored in the metadata that follow them.

The paper from the EMP published in 2017 gathers 27,751 samples from 97 independent studies. The collection of these material extracts is standardized relative to the nature of the "world" to be sampled. For example, material extracts are constituted by swabs for microbial communities found on surfaces. For water communities, material extracts are filters; for communities of soil, sediment, and feces, material extracts are bulk samples. They "were collected fresh and, where possible, immediately frozen in liquid nitrogen and stored at -80 °C" (Thompson et al. 2017, Supplementary Methods). Moreover, the project created a specific bio-ontology[5] (EMPO). EMPO assigns samples to environments according to different criteria (e.g., host-associated or free-living microbial communities) (Thompson et al. 2017).

[5] Leonelli 2016 discusses the role of these bio-ontologies in data traveling and reuse (see Chapter 5 and section 5.2). I will only state that if EMP ontology facilitates the reuse of the computerized DNA sequences, it also limits the number of situations in which these data can travel by fixing certain parameters and their relations with each other.

Step B – Molecular DNA is extracted from these material extracts after several biochemical reactions, including microorganismal membrane disruption. Hence, the DNA of a particular microorganism is mixed with DNA extracted from other occurring microorganisms. As a result, the DNA-microorganism link is lost. This molecular DNA is stored to be used for amplification.

Step C – In the AS method, a specific DNA part, the amplicon, is amplified through biochemical reactions. When studying microbial communities, two assumptions hold. This is true regardless of the scientific question at hand. First, the amplicon represents the whole microorganism. It can thus reconstruct the missing link between the DNA and the microorganism. Second, the amplicon sequence is evolutionarily stable and novel enough to separate microbial species and reliably measure the natural diversity of the microbial community. Thus, these molecular DNA amplicons are the material basis for future knowledge claims, and parts of them are sent to sequencing facilities.

Material extracts, molecular DNA, and molecular amplicons are called “samples” by scientists. These products are all stored frozen and handled so that they can be moved between different locations. They are handled to enable their dissemination. In theory, they can be used in new scientific inquiries, but in practice, they are usually discarded after being processed in a single inquiry. Scientists treat them as containing the information that will act as evidence in the scientific inquiry. However, these scientific products must be transformed before fulfilling this role. In contrast, “controls” in AS are very similar in nature but not in function to

"samples." Scientists consider "controls" as elements that can inform the reliability of the experimental process but not as containing evidence for knowledge claims. In alignment with the relational framework, this example shows that two similar objects are distinguished not on the basis of essential differences but on the basis of the situation of inquiry and the researchers within it.

These samples are the initial link to the external world. Information about them (metadata) grounds the continuity of the reference between the external world—the environment—and the knowledge about it—this environment is composed of this microbial community. Samples and their metadata are why the other kinds of data produced afterward are still about something in the world. The samples themselves are lost; if their metadata are also lost, the computerized DNA sequences produced (step D) cannot refer to or represent the outer material world.

Step D – Sequencing facilities have specific equipment and human skills. They sequence the amplicons, resulting in computerized DNA sequences. The molecular information of the nucleotide chain is converted into strings of letters. These strings can be manipulated and compared computationally. However, they have no biological meaning without extra information. Scientists call these computerized DNA sequences "raw data." The computerized DNA sequences are deposited in databases with technical metadata. They travel, for example, from the sequencing facility's computer to the researcher's computer. This is already a change in the situation of scientific inquiry. The objectives are different: technical, i.e., to get a

reliable object for constructing knowledge claims during sequencing; biological, when these sequences are used to describe a community, to explain the microbial distribution, etc.

In the EMP, steps B to D are standardized to facilitate the reuse of the computerized DNA sequences (Pollock et al. 2018). In particular, the authors insist on three critical points for these steps: the use of a single protocol for step B (DNA extraction), which is chosen for its efficiency toward diverse sample types, rather than its high efficiency toward any given type. The primers for step C (Amplicon amplification) are the same across individual studies and are chosen with the abovementioned assumptions in mind. Finally, the sequencing technologies are limited to Illumina HiSeq and Illumina MiSeq (Thompson et al. 2017, Supplementary Methods).

The "raw" data travel with their metadata. These metadata register the detailed circumstances of producing these data in each individual study of the EMP project. Their quality was assessed before integrating the pool of studies used in the 2017 publication (Thompson et al. 2017).

Step E – Computerized DNA sequences are transformed through many computational and statistical steps. The underlying assumption is that statistical methods allow scientists to distinguish between natural and artifactual variation (caused by the AS method itself) in DNA. To date, there are no community-accepted standards (Pollock et al. 2018). However, such decisions affect the knowledge claims made later. This step leads to what I call "pre-treated DNA sequences." These sequences are stored locally (in the researcher's computer) and can be

disseminated in principle. However, this is rarely the case because of the lack of standardization. By modifying computerized DNA sequences, the scope of what they can represent is narrowed down to a single organism or type of organism. For example, each computerized DNA sequence contains a small part at the beginning. This region corresponds to the initiation of the DNA amplification and sequencing procedures. It can be removed or kept. If it is kept, further analysis can be performed on this sequence, such as to determine how far it is genetically conserved. If not, the sequence cannot represent genetic variations in this region.

Step F – Pre-treated DNA sequences are sorted according to their variations. The distinction between natural and artifactual variations is related to the assumption of the method. Given the range of mathematical and statistical tools available, this method impacts the inquiry's conclusions. This step is called "denoising" or "clustering," depending on the method used. "Denoising" refers to the elimination of errors—the artifactual variations. Instead, "clustering" refers to the grouping of sequences by similarity. Sometimes, data are cleaned to be more suitable for travel (Boumans and Leonelli 2020). Here, the sequences are cleaned to ensure the reliability of the products they will help construct. These products will ground knowledge claims about a particular environment's microbial composition and diversity. The sequences are cleaned to be able to sustain knowledge claims.

This step leads to the obtention of occurrence tables (Figure 1). The rows are the amplicon sequence variants (ASVs). These are DNA sequences of the region amplified (cleaned pre-

treated DNA sequences). The columns correspond to the material extracts. Each cell contains the number of copies of an ASV retrieved in a material extract. The aim of this step is twofold. First, to count how many different individual microorganisms (i.e., different ASVs or the number of rows) the original environment might contain. Second, to count how many of these individuals are detected relative to the total number of microorganisms retrieved. The ASVs now represent only one kind of thing, a specific kind of microorganism (depending on the resolution, it can be a species or a genus). This representational image is fixed within the situation of inquiry that has produced it. In community ecology, occurrence tables play the same role as observational diversity measurements do. Once again, the AS method is a molecular alternative to microscopic observation.

Occurrence tables travel between different situations of inquiry in the form of published papers (usually in the supplementary information). However, they are not reused in different situations of inquiry.

In the EMP example, occurrence tables are called “observation tables” (Thompson et al. 2017, Supplementary Methods). Each original study generated occurrence tables for its line of investigation using the computerized DNA sequences it obtained. These tables are (most of the time) accessible in the supplementary data of these individual studies. Each separate study is a data production and analysis situation that generates “raw data” (computerized DNA sequences) and occurrence tables. Thompson and coauthors then developed new tables to uncover patterns of microbial diversity globally. When gathered for the EMP project, these

computerized DNA sequences travel into a new situation of analysis, but not the occurrence tables.

Step G – Occurrence tables are combined with metadata and knowledge stored in databases (existing taxonomy) to serve as evidence for knowledge claims about the microbial community. Scientific papers constitute knowledge claims or at least contain them; they are published and travel between different situations of inquiry.

The EMP project has a double output: one about the physical world and one about the inquiry process. Biological conclusions involve the determination of patterns in community distribution and the exploration of "key hypotheses in ecological theory" (Thompson et al. 2017, 458). The authors also highlight the creation of a reference database and a standardized framework for incorporating data from future studies as achievements of their work.

3.2. The Shortcomings of the Relational Framework

In the AS case, there is causal continuity between all the empirical products. After all, the material extracts contain the DNA that will end up populating the occurrence tables. These findings will then ground conclusions about the microbial community. Scientists consider these products to be potential evidence. Moreover, given the understanding of travel developed in section 2.2, all these products have the potential to travel. All these products can be stored locally in virtual or physical databases, and they can all be distributed by mail or

email to researchers for analysis in various investigations. Thus, all the AS products fulfill conditions 1) and 2) of the relational view; they are all data (Table 1).

This inclusivity of the relational framework helps make sense of the diversity of objects used as data in scientific inquiry. However, it does not help to understand the distinctive role played by samples compared with data in the AS case and, more generally, biological research.

Scientists refer to some objects of the inquiry as "samples" and others as "data." As philosophers, we should be cautious when analyzing these concepts before equating the philosophical ones with those used by scientists. I argue that we should at least consider this difference in the discourse relevant.

First, the relational framework is aligned with the practice-oriented tradition. Indeed, one of the advantages of the relational framework is its relevance to biological scientific practice and how it deals with data. The language used by scientists is one of these practices and thus should be considered when analyzing scientific objects.

Second, despite the lack of homogeneity in the use and meaning of "samples" and "data" in the scientific discourse (as is the case with words such as gene or species), here, I indicate that the scientific use of different words highlights the need for different concepts designating different things in the inquiry. I propose that this distinction is, at least partly, based on the distinctive role of these two objects in the inquiry. As described above, molecular DNA does

not play the same role as computerized DNA sequences. The evidential role of what scientists call samples and what they call data (raw or not) is different.

More generally, as highlighted in the introduction, the analysis of samples as scientific objects could be more precise (Leonelli and Tempini 2020, viii and 17). The connection between samples and data remains to be clarified to characterize the full extent of biological scientific practice epistemically.

This paper contributes to the issue of the place and epistemic role of samples in the inquiry. I propose precisions of the criteria of the relational framework that help distinguish the peculiar position of samples compared with other kinds of data in the case of AS (Section 4). These specifications are compatible with the existing literature on data and evidence (Section 5). Because this ambiguity might be due to discipline-related differences, I propose a first step toward a generalization in the conclusion.

Table 1

Mapping of the products of amplicon sequencing with the conditions of the relational framework

Step. Products of AS inquiry	Considered as potential evidence	Capacity to travel between different situations of inquiry	Relational framework
A. Material extracts	Y	Y	Data
B. Molecular DNA	Y	Y	
C. Molecular DNA amplicons	Y	Y	
D. Computerized DNA sequences	Y	Y	
E. Pre-treated DNA sequences	Y	Y	
F. Occurrence tables	Y	Y	
G. Scientific papers	N*	Y	Knowledge claims

*I take them to contain the evidence either explicitly or implicitly.

4. Samples

In AS and within the relational framework, samples 1) are *necessarily transformed* to be considered evidence, 2) are portable *in the limit of the situation where they endure this transformation,* and 3) *act as material/world anchors for claims about a phenomenon.* I clarify these points successively in the remainder of the section.

Samples are a subcategory of data because they fulfill the criterion of being treated as *potential* evidence, but they must be transformed to be considered evidence. In the general definition of data from the relational framework, data are considered *potential* evidence. The scientific product that will end up being considered evidence is a subset of the ones considered *potential* evidence. I argue here that samples are not part of this subset. They are treated and considered *potential* evidence. If not, there will be no difference between a bulk of soil collected for no purpose and a bulk of soil collected in the context of a scientific inquiry. In addition, samples do not need to be transformed to be considered *potential* evidence. They are collected from the world according to an inquiry procedure. However, samples differ from other kinds of data because they can never be evidence per se. They need to be transformed, changed, and altered. This transformation is necessary. In contrast, the result of step D, the computerized DNA sequences, could readily be used as evidence. These "raw" data are extracted from the samples. This form and other data are not automatically evidence; it will depend on the situation of inquiry and often on additional information, but they have the potential to do so; transformation is not necessary.

The necessity of the transformation comes from the fact that the samples are not amenable to human understanding. An investigator can do nothing with a hand of soil; she needs to extract something from it. This extraction or transformation needs to lead to something amenable to human understanding, manipulation, and the starting point of inferential reasoning. Transforming samples into something about the world, another form of data, is a transition from "being part of the world" to not being a part of it anymore. In AS, molecular sequences are transformed into letter sequences that refer to the world because of the causal continuity between the samples and DNA sequences.

I parallel the notion of samples with that of "ubjects" to emphasize the importance of the situation and the inquirers in what constitutes a sample. In the context of exchanging human body material, the concept of a "ubject" has been developed as a hybrid between an object and a subject: "Ubject [designates] that which is seen by (at least) some as having been part of a body and therefore related to a subject." (Hoeyer 2013, 5). "Ubjects" are those things leaving "the space identified as the body" (Hoeyer 2013, 68). They should not be conceived "as an entity with a name but more like a point in time between being part and not part of a body" (Hoeyer 2013, 143)[6]. Similarly, samples are those entities seen by some inquirers as leaving the space identified as the world. The sample is a point in the inquiry between being part and not part of the world. They are a hybrid between a portion of the natural world and a

[6] I thank Sara Green for bringing this work to my attention.

representation of it, a “presentation.” For example, the bulk of soil sampled in step A is a portion of the world, and at the same time, it represents the environment of this portion of the world. I return to the samples’ representational power later in this section.

A rupture of the material continuity of the scientific product marks the transformation of samples into another kind of data. In the general definition of data, rupture of material continuity does not automatically include or exclude a scientific object in the set of objects considered data. The nature of the object does not matter fundamentally. Here, I emphasize this change to highlight that samples are a distinct kind of data and thus need particular analysis. It highlights the scientific practice of sampling and sample processing and allows a focus on these “upstream” activities rather than data analysis. In AS, the materiality of the sample is essential for the epistemic relevance of the DNA sequence retrieved. There is an “integration of physical matter from various sources so as to constitute a new entity” (Halfmann 2020, 27). Indeed, after DNA extraction – DNA comes from the “world” – nucleotides and other things are added to synthesize new DNA. This step integrates new physical matter. The physical part coming from the world might be lost at the end of the process. What remains is the information on the organized sequence of nucleotides. So, the material continuity of the samples (steps A to C) is conserved. The sample is preserved “through several stages of the research process without a change of medium” (Halfmann 2020, 27). In AS, between steps A and C, “information is transferred materially and not by impression or translation to a different medium.” (Halfmann 2020, 39). The transformation

from samples to step D is a “jump” from one medium to another; it is the digitalization of the DNA sequence, a rupture of the material continuity. It is not a degree of transformation that distinguishes samples from other kinds of data but rather the necessity of the transformation.

In the relational framework, data are portable. They have the potential to travel, and they are handled to travel between different situations. In AS, samples are handled to travel between different facilities, at least from the field to the laboratory (Figure 1). They are portable. The potential to travel between situations does not determine whether the object will travel. Scientific products travel between situations, or they do not. I argue that samples travel peculiarly because they change location and time, but this is not a change of situation. The research question is the same: identifying patterns in the microorganismal population of a particular environment (a classic question in community ecology). The relevant criteria of the context are the same during steps B and C in the laboratory as they are during the collection of samples, i.e., step A in the field. Thus, samples travel within the same situation between different phases of the inquiry but not between different situations. This is important in AS because it is a fundamental difference between the “samples” and other products of the inquiry.

Another way to put it is to use the distinction between “phenomena-time” and “data-time” (Leonelli 2018) when data are limited to the kind that I am delineating: samples. In AS, the time and location of the samples are essential to the phenomenon under investigation. They constitute information of the “phenomena-time.” They are relevant information gathered in the

metadata file. They always accompany other kinds of data extracted from those samples. However, the time and location of the sample transformation are contextual information that is not relevant to the investigation and thus does not pertain to the situation of inquiry. The methods used are relevant, so they are part of the situation and of the "data-time". They influence the evaluation of the quality of the data retrieved and their future uses. This is compatible with Leonelli's vision of the situation. Data can travel between their production site and utilization site. These sites might constitute different situations, but not necessarily. I argue that in AS, samples are a kind of data that stays in one situation. They already become something new when they change situations (for example, between steps C and D). They are in transition to become another kind of data.

Another important specification for samples in AS inquiry is that they are portable but only portable once. Indeed, samples are discarded once computerized DNA sequences are obtained. In AS inquiry, sample analysis is never reproducible. It is physically impossible to reanalyze the same set of samples in the same manner. It is a disruptive science. The destruction of research products tracks a distinction between samples and other kinds of data. Research products might be destroyed, and only the latter stage of the "data process" might be conserved. For example, not all the "intermediate" files generated during computerized DNA sequence analysis are conserved; they are destroyed. The status of any given research product—data or not—will depend on whether it has been handled to travel. In AS, samples are handled in ways that enable their circulation. Moreover, the samples are destroyed, but

information about them (metadata) is not. As the EMP example suggests, the computerized DNA sequences are mobile and available for reanalysis, but this is not the case for the samples.

Samples are replicable but nonreusable. That is, samples are renewable in that an experimenter can, in theory, return to the experimental field, take another sample of the same environment as the first one, and extract what she might consider similar "raw" data. However, the original sample is lost forever. Moreover, in that case, the "phenomena-time" is different from the original one, which might have an impact depending on the research question. Thus, the sample is replicable but nonreusable.

Samples have dual representational content. On the one hand, samples are literally part of the world. They are the products of material continuity. There is no intentional "writing or imprinting information onto a medium" (Halfmann 2020, 39, footnote 22). Samples are representative instances of the target; in this sense, they have a similar representative role as that of the statistical samples. The idea is that an investigator can take whichever part of a homogenous environment. Samples are subsets, typical, characteristic, or usual of a more extensive set, the environment, that is investigated. However, this representative role is nonintentional and contingent on the sample being the environment. This does not mean that it is fixed and immutable. The material process itself threatens this fixity. For example, between the field and the laboratory, samples are frozen to try to keep their microorganismal

composition fixed. Different techniques of freezing impact the composition retrieved later (Pollock et al. 2018, 2).

Samples speak about the world because they are the world as it is. If the samples have been correctly collected, there is no question about the appropriateness or accuracy of this representation in the same sense as in whether the data accurately represent a phenomenon. The question is how representative the sample is of the whole but not how samples represent the whole. That is, whether samples are typical and relatively preserved instances of the phenomenon. Samples are anchors for data claims; they ensure that these data document this phenomenon (Wylie 2020)[7]. In AS, samples are why researchers can say that their strings of letters speak about the world. Samples are the physical reference of propositions written in a lab book. In a Latourian sense, I consider samples to be the initial step of a reference circle. Samples play the role of the referent that scientists can "point to with their finger outside of discourse," even if it needs to be transformed to be "brought back inside discourse" (Latour 1999, 32). The difference with Latour's case study is that the reference is lost in AS: the samples are destroyed. However, the information about them is conserved.

[7] Wylie also considers samples as data. The word "data" in her work refers to two things. It refers to the final product of a "practical argument," i.e., the "evidential anchor." It also refers to a part of this argumentation, i.e., the samples but also the "radiogenic data," the "temporal data," and the "chronological data."

On the other hand, samples instantiate the properties of the whole environment(s) studied. A scientific intention is needed to interpret these properties. Again, this is another difference between a soil sample and a handful of soil. In AS, samples instantiate two kinds of relevant properties. The properties, such as the temperature, pH, time, date, etc., of the samples were recorded as metadata. These parameters affect the environment that scientists seek to understand. The other properties, such as the microorganismal composition and organization of the sample, are the ones under study. The first properties determine which kind of environment the sample is from. They have a double use of characterizing the environment of interest, for example, "dry land," and monitor that different samples of the same environment are similar enough to pertain to the same category. Samples are "typical" enough of the environment. They do not display exceptional values for those parameters. The typicality of these parameters is crucial for investigation because it warrants inference according to which other properties, such as the microorganismal composition, are also typical of the environment studied. This will then warrant inferences that data extracted from these samples document the entire environment and not only the tokens of that environment, i.e., the samples. The products from steps A and C are primarily handled to secure the typicality of the DNA in terms of identification (the DNA is typical of this or that species) and composition (there is molecular DNA of all the species and only the species that are present in the original environment and the same relative quantity). This is why samples are usually discarded after sequencing. They cannot be reused because they are no longer considered typical for these parameters (DNA can

have been modified or lost). This loss of typicality will lead to impairments and less reliability in the data extracted, ultimately threatening the claims based on these data.

I want to make two remarks. The first concerns the compatibility of this conceptualization of samples with the relational framework, and the second concerns the connection of this analysis with the notion of specimenhood[8] (Currie & Levy 2019).

The relevant properties of similar samples can differ depending on the situation. For example, in a biological context, the relevant properties of a bulk of soil might be how it has been processed and the methods used. The bulk of soil then instantiates those properties, relegating the microorganismal composition to the background of the investigation. Similar or different samples of the same environment can instantiate other properties according to the situation. It is compatible with the relational framework of data that attributes a representational scope to data that can change between situations of scientific inquiry (Leonelli 2019).

The notions used to characterize samples' representational content and role in AS are close to the idea of specimenhood[9] developed in the context of experiments (Currie & Levy 2019). Like specimens, samples in AS are "part of the world [brought] into the lab" (Currie and Levy 2019, 1071). They are objects typical of a target, the environment, and they are drawn by an

[8] I thank Franziska Reinhard for bringing this literature to my attention.

[9] This concept is close to the one of exemplification developed by Elgin – see footnote 4 for an explanation of the difference between the two (Currie and Levy 2019, 1073).

"unbiased procedure" or at least a procedure that "reduces the risk of selecting an unusual object, and which preserves typicality by avoiding problematic alteration of the object" (Currie and Levy 2019, 1073). Whether the AS procedure causes "problematic alteration" does not fall within the scope of this paper, and instead, I focus on two differences between specimens and samples. First, the difference between the object and the target is not empirically tractable. In AS, there is no way of precisely knowing the distance in the "focal property," here the microorganismal composition, between the sample and the environment studied. The best scientists can do is multiply the number of samples, standardize the collection protocol, and use various controls. However, ultimately, what scientists analyze is only a picture of the real world. Second, the notion of specimenhood is developed in the context of hypothesis-driven experiments. In contrast, AS is closer to observation. The occurrence table obtained during the inquiry is an adaptation of the observational fieldwork in community ecology. The methodological principle is the same: counting the number of individuals per species in a restricted area. This is done in macroecology. In microecology, this has been done using a microscope. AS is like looking with a microscope at who is there, how many they are, and how they are distributed within the environment. The aim is the same: describing the ecology of the community and identifying new phenomena, i.e., patterns of diversity that will have to be explained.

5. Branching out of the Relational Framework

5.1. Other Kinds of Data

No data are absolutely unprocessed. However, the AS case analysis shows a difference in the role of three kinds of data in the inquiry: samples (detailed above), "raw" data, which are relatively unprocessed, and non-raw data, which are relatively more processed.

They can be interpreted as different stages of the data journeys. Similar to the case described in Tempini (2020), computerized DNA sequences – "raw" data – are "data sources." These sequences can be selected, reassembled, and transformed into "data mixes." One can generate as many different data mixtures of combinations and modifications of the data sources as possible. However, once you have settled your mixes in a new table, they are more fixed than the data sources used to generate them. Trade-offs are needed in this transformation or "metamorphosis" (Tempini 2020, 259). In AS, the key trade-off between computerized DNA sequences and occurrence tables is determining the boundary between natural and artifactual variation in DNA sequences.

Transforming biological information into digital data is likened to a "pipeline" (Stevens 2013, 109)[10]. This metaphor applies to AS, where "raw" data are converted into non-raw data through "bioinformatic pipelines." The distinction between "raw" and non-raw data is relative.

[10] Stevens uses this metaphor for sequencing platforms. This is similar to steps C to D in the current analysis, albeit the sequences are not put into GenBank by the sequencing platforms, the scientists upload them into databases. Additionally, only one part of a gene (not even necessarily the entire gene) is sequenced – the "amplicon" – not whole genomes.

They can be distinguished within the virtual space (Stevens 2013, 129-130). In AS, "raw" data – computerized DNA sequences – are publicly available in databases[11], standardized under digital formats (e.g., FASTQ) and bio-ontologies, and preserved for future uses. In contrast, non-raw data are locally accessible and modifiable only to the bioinformatician who produced them. Bioinformaticians are crucial in making data usable for biological research and knowledge production (Stevens 2013, 127-133). In AS, the uploading of "raw" data into databases and the transformation of "raw" data into non-raw data can be performed by the same person, be it a biologist with bioinformatics training or a bioinformatician with "wet" lab training, exemplifying a hybrid epistemic practice described by Strasser as something between "collecting" and "experimenting" (2019, 108, 115). This process influences how biological information is perceived and ultimately shapes the form of knowledge produced, with standardization of steps within the pipeline affecting the definition of "raw" data. For example, if steps D to E become standardized in AS, the "raw" data would be the pre-treated DNA sequences instead of the computerized DNA sequences[12].

[11] For a detailed analysis of the role of databases and other collection of specimens and data in the production of scientific knowledge, see Strasser (2019).

[12] I am grateful to an anonymous reviewer for highlighting the importance of this literature in the context of this analysis.

Another way to analyze the difference between "raw" and non-raw data is to consider the occurrence tables (Figure 1, result of step F) as data models. The relational framework sees data models as the "ordering of data" to represent a more specific phenomenon or to make one or more patterns salient in a collection of data (Leonelli 2019). Indeed, occurrence tables are partly created by sorting the pre-treated DNA sequences. In constructing them, researchers want to make the natural diversity salient compared to the diversity generated via the AS method. I have two comments on this point. First, scientists do not consider occurrence tables to be models. They are reluctant to do so because they do not consider the mathematical and statistical tools used to get the occurrence tables to model anything (although some of these tools may use models to do their work). As mentioned above, these tables are considered more as observations, not as a model of these observations. Second, if we put aside the last comment and assume that they are data models in the relational sense, that would not undermine the conclusions drawn for the categories of samples and the difference in their role compared with the role of other kinds of data. The generalizability of these precisions to different use cases in the life sciences remains to be established.

5.2. Enriched Evidence

I draw a parallel with the notion of "enriched evidence" (Boyd 2018). She developed the idea that theories should be tested against enriched evidence to at least meet the empirical criterion of empirical adequacy. Enriched evidence contains a kind of empirical constraint that the theory needs to meet to remain relevant in current scientific practice. Enriched evidence is

defined as follows: “The evidence with respect to which empirical adequacy is to be adjudicated is made up of lines of evidence enriched by auxiliary information about how these lines were generated.” More precisely, a line of evidence is “a sequence of empirical results including the records of data collection and all subsequent products of data processing generated on the way to some final empirical constraint.” Auxiliary information is “the metadata regarding the provenance of the data records and the processing workflow that transforms them” (Boyd 2018, 406-407).

In this analysis, I take samples to be at the origin of the “line of evidence.” They are the first empirical results that scientists take as valuable to distinguish because they serve to produce and construct the data. They are not constraining enough; data need to be extracted. In this framework, thus, the samples described above would be different from the data conceptualized here.

The difference between the data kinds mentioned in section 5.1. could be interpreted as a difference between an empirical result and an empirical constraint. Data and auxiliary information are part of the final “empirical constraint” (Boyd 2018). They ground knowledge claims by constraining the interpretation that can be done using them. However, in my case study, the final empirical constraint is the association of the occurrence table, the metadata associated with each sample, and the taxonomy retrieved using the DNA sequences and information stored in databases (see Figure 1). I would instead consider the computerized DNA sequences as something “mal-adapted” (Boyd 2018) to a theory because they are not

usually produced to test a theory. Therefore, they need to be transformed to be better adapted (and become occurrence tables) to describe a phenomenon. Contrary to Boyd, I do not think that data always test theory. They can be very descriptive or explorative. Nevertheless, they empirically constrain what theory or explanations scientists choose to follow or deepen.

5.3. The Pragmatic-Representational View of Data.

I draw a parallel with the "pragmatic-representational (PR) view of data" (Bokulich and Parker 2021). They place representation at the center of the data's role while embracing the view that all data are constructed through a process of inquiry. That is, "they should be evaluated in terms of their adequacy or fitness for particular purposes" (Bokulich and Parker 2021, 1). Like Leonelli, constructed does not equate to data having content entirely subjective to scientists' choices. They also recognize that data have a sort of representational "scope" dependent on the world, the scientists, and the various elements of the situation in which the inquiry occurs.

Their view aims to account for how data and data models are about the world, "no matter how many rungs we have in our data model hierarchy, at some point we need our ladder to reach the ground" (Bokulich and Parker 2021, 5). I argue that because samples are part of the world and not only representations of it, they can be the first step of that ladder. Samples are the world anchors for data claims about the phenomenon. This approach is compatible with the idea that the extent to which any given sample, data, or dataset can still document the world should be assessed in each situation of inquiry.

The idea of a jump from one medium to another medium between samples and other kinds of data aligns with the idea of the PR view that what counts as data is narrower than in the relational framework. Indeed, only records of the interaction process between the world, an apparatus, and an inquirer count as data (Bokulich and Parker 2021, 6-7). Moreover, in this view, the objects I described to be sampled will not be data as in the relational framework but part of the phenomenon itself. I grant that point in my precision of what samples are (Section 4), yet because they are also considered epistemically as having the capacity to contain data, they are more than that. Thus, samples are data in the sense developed by Leonelli (Table 1) but are a particular and distinctive kind of data.

The difference from the PR view is that samples should not be judged according to "whether they can be used to achieve the particular epistemic or practical aims that interest their users" (Bokulich and Parker 2021, 10). The question in sampling is different, and the epistemic aim might be in the background of the sampling method. What is essential about samples is how well they retain the typicality of the phenomenon (or some properties of it) they are an instance of.

A second difference is that samples cannot be reused in the sense developed in the PR view. "While data reuse involves using the same data to answer the same question, data repurposing involves using the same data to answer a different question" (Bokulich and Parker 2021, 11, original emphasis removed). Their definition of data reuse is close to the notion of reproducibility explored in section 4. In this sense, samples are not reusable, as they are

usually discarded. Investigators need to generate new samples or take computerized DNA sequences.

In the EMP example, the justification for the study's repeatability is that the computerized DNA sequences come from multiple samples from multiple studies for each habitat, constituting a replication but not a reproduction. The EMP is an example of the repurposing of computerized DNA sequences. As argued above, a new analysis of the same data is a situation change from the first analysis, which generated those data. The question remains as to whether the situation of inquiry needs to change in order to repurpose data. It is not my aim here to answer such a question.

6. Conclusion and Outlook

I propose refinements to the criteria of the relational framework. These specifications help understand samples' distinctive role in scientific inquiry compared with other kinds of data in the case of AS. By being compatible with additional views of data and evidence, these specifications can help analyze other life sciences cases.

As a first step, I explore the case of biobanks and argue that these modifications can help understand the epistemic role of different scientific objects. Biobanks are infrastructures that collect, manage, and store vast quantities of biological materials (human or not). They also deliver these materials to different research groups.

The term “sample” is used to designate the biological extract at the beginning of the biobanking process and the biological material delivered to a research group. Thus, it would seem that my criterion of “traveling only within the situation of data production” does not apply. While I agree that this constitutes two different situations, I argue that a particular object can be considered as sample in one situation and as another kind of data in another—aligning with the relational framework—depending on the inquirer and the situation at stake.

In the situation of biobanking, a biological extract (e.g., blood) is taken. It is moved from the collection site to the biobank site. This extract is necessarily transformed (treated and processed) before being considered evidence. Several tubes (e.g., different pretreatments before storage) can be made with the same biological extract. All these tubes are stored in a collection managed and organized by biobankers. Contextual information about the origins of the extract, how it has been processed, etc., is also stored in a digital format, i.e., metadata. The biological extract is the material anchor for different technical or scientific claims. So, these biological extracts are samples, whereas the tubes preserved in biobanks are “biological resources” or even “material data” (Clarizio 2022).

When materials are sent or requested for a particular inquiry, they reach a new situation of inquiry. Inquirers evaluate the role of the scientific objects at their disposal. When used in these cases, these “biological resources” play the role of samples. They necessitate transformation before being able to be considered evidence. They will move, even between different laboratories, to produce this evidence. Inquirers use these materials (and not the

biological extract they are from) as anchors to ensure that downstream data document a phenomenon in the world.

In this paper, I provide specifications of the relational framework that help analyze the distinctive role of samples compared with data in scientific inquiry. These specifications help deepen our understanding of these objects and their epistemological role.

7. References

Alteio, Lauren v., Joana Séneca, Alberto Canarini, Roey Angel, Jan Jansa, Ksenia Guseva, Christina Kaiser, Andreas Richter, and Hannes Schmidt (2021) A Critical Perspective on Interpreting Amplicon Sequencing Data in Soil Ecological Research. Soil Biology and Biochemistry 160. https://doi.org/10.1016/j.soilbio.2021.108357.

Bokulich, Alisa, and Wendy Parker (2021) Data Models, Representation, and Adequacy-for-Purpose. European Journal for Philosophy of Science 11 (31). https://doi.org/https://doi.org/10.1007/s13194-020-00345-2.

Boumans, Marcel, and Sabina Leonelli (2020) From Dirty Data to Tidy Facts: Clustering Practices in Plant Phenomics and Business Cycles Analysis. In: Sabina Leonelli and Niccolò Tempini (eds) Data Journeys in the Sciences. Springer, Cham, pp 79–101. https://doi-org.libproxy.viko.lt/10.1007/978-3-030-37177-7_5.

Boyd, Nora Mills (2018) Evidence Enriched. Philosophy of Science 85 (3): 403–21. https://doi.org/10.1086/697747.

Boyd, Nora Mills, and James Bogen (2021) Theory and Observation in Science. In: Edward N. Zalta (ed) Stanford Encyclopedia of Philosophy, Winter 2021 Edition. https://plato.stanford.edu/archives/win2021/entries/science-theory-observation/.

Brown, Matthew J (2012) John Dewey's Logic of Science. HOPOS: The Journal of the International Society for the History of Philosophy of Science 2 (2): 258–306. https://doi.org/10.1086/666843

Clarizio, Emanuele (2022) La Production de La Valeur Épistémique Des Ressources Biologiques Dans Les Biobanques. In: Emanuele Clarizio, Céline Chérici, Jean-Claude Dupont, Xavier Guchet, and Yves-Édouard Herpe (eds) Conserver Le Vivant : Les Biobanques Face Au Défi de La Médecine Personnalisée. Editions Matériologiques, pp 119–34.

Currie, Adrian (2021) Stepping Forwards by Looking Back: Underdetermination, Epistemic Scarcity and Legacy Data. Perspectives on Science 29 (1): 104–32. https://doi.org/10.1162/posc_a_00362.

Currie, Adrian, and Arnon Levy (2019) Why Experiments Matter. Inquiry (United Kingdom) 62 (9–10): 1066–90. https://doi.org/10.1080/0020174X.2018.1533883.

Dewey, John (1938) Logic: The Theory of Inquiry. Holt, Rinehart, and Winston, New York.

Griesemer, James (2020) A Data Journey Through Dataset-Centric Population Genomics. In: Sabina Leonelli and Niccolò Tempini (eds) Data Journeys in the Sciences. Springer, Cham, pp 145–67. https://doi.org/https://doi.org/10.1007/978-3-030-37177-7_8.

Halfmann, Gregor (2020) Material Origins of a Data Journey in Ocean Science: How Sampling and Scaffolding Shape Data Practices. In: Sabina Leonelli and Niccolò

Tempini (eds) Data Journeys in the Sciences. Springer, Cham, pp 27–44. https://doi.org/10.1007/978-3-030-37177-7_2.

Hoeyer, Klaus (2013) Exchanging Human Bodily Material: Rethinking Bodies and Markets. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-5264-1.

Howlett, Peter, and Mary S. Morgan, eds. (2011) How Well Do Facts Travel? The Dissemination of Reliable Knowledge. Cambridge University Press, New York.

Latour, Bruno (1999) Pandora's Hope: Essays on the Reality of Science Studies. Harvard University Press, Cambridge.

Leonelli, Sabina (2016) Data-Centric Biology: A Philosophical Study. The University of Chicago Press, Chicago and London.

Leonelli, Sabina (2018) The Time of Data: Timescales of Data Use in the Life Sciences. Philosophy of Science 85 (5): 741–54. https://doi.org/10.1086/699699.

Leonelli, Sabina (2019) What Distinguishes Data from Models? European Journal for Philosophy of Science 9 (22): 1–28. https://doi.org/10.1007/s13194-018-0246-0.

Leonelli, Sabina, and Niccolò Tempini eds. (2020) Data Journeys in the Sciences. Data Journeys in the Sciences. Springer, Cham. https://doi.org/10.1007/978-3-030-37177-7.

Lloyd, Elisabeth, Greg Lusk, Stuart Gluck, and Seth McGinnis (2022) Varieties of Data-Centric Science: Regional Climate Modeling and Model Organism Research. Philosophy of Science 89 (4): 802–23. https://doi.org/10.1017/psa.2021.50.

Mansnerus, Erika, and Susann Wagenknecht (2015) Feeling with the Organism: A Blueprint for an Empirical Philosophy of Science. In: Suann Wagenknecht, Nancy J. Nersessian and Hanne Andersen (eds) Empirical Philosophy of Science, Introducing Qualitative Methods into Philosophy of Science. Springer, Cham, pp 37–61. https://doi.org/10.1007/978-3-319-18600-9_3.

Morgan, Mary S (2011) Introduction. In: Peter Howlett and Mary S. Morgan (eds) How Well Do Facts Travel? The Dissemination of Reliable Knowledge. Cambridge University Press, New York, pp 3-42.

Peschard, Isabelle F., and Bas C. van Fraassen (2018) Introduction. In: Isabelle F. Peschard and Bas C. van Fraassen (eds) The Experimental Side of Modeling. The University of Minnesota, Minneapolis, pp 1-58.

Pietsch, Wolfgang (2015) Aspects of Theory-Ladenness in Data-Intensive Science. Philosophy of Science 82 (5): 905–16. https://doi.org/10.1086/683328.

Pollock, Jolinda, Laura Glendinning, Trong Wisedchanwet, and Mick Watson (2018) The Madness of Microbiome: Attempting To Find Consensus ‘Best Practice’ for 16S Microbiome Studies. Applied and Environmental Microbiology. 84 (7): 1–12.

Romeijn, Jan-Willem (2022) Philosophy of Statistics. In: Edward N. Zalta and Uri Nodelman (eds) The Stanford Encyclopedia of Philosophy, Fall 2022 Edition. https://plato.stanford.edu/archives/fall2022/entries/statistics/

Soler, Léna, Sjoerd Zwart, Vincent Israel-Jost, and Michael Lynch (2014) Introduction. In: Léna Soler, Sjoerd Zwart, Vincent Israel-Jost, and Michael Lynch (eds) Science After the Practice Turn in the Philosophy, History, and Social Studies of Science. Taylor & Francis, New York, pp 1–43.

Stevens, H (2013) Life Out of Sequence: A Data-Driven History of Bioinformatics. The University of Chicago Press, Chicago and London.

Strasser, B. J. (2019) Collecting Experiments: Making Big Data Biology. The University of Chicago Press, Chicago and London.

Suppes, Patrick (1962) Models of Data. In: Ernst Nagel, Patrick Suppes, and Alfred Tarski (eds) Logic, Methodology and Philosophy of Science. Stanford University Press, Stanford, pp 252–61.

Tempini, Niccolò (2020) The Reuse of Digital Computer Data: Transformation, Recombination and Generation of Data Mixes in Big Data Science. In: Sabina Leonelli and Niccolò Tempini (eds) Data Journeys in the Sciences. Springer, Cham, pp 239–63. https://doi.org/10.1007/978-3-030-37177-7_13.

Thompson, Luke R., Jon G. Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J. Locey, Robert J. Prill, et al. (2017) A Communal Catalogue Reveals Earth's Multiscale Microbial Diversity. Nature 551 (7681): 457–63. https://doi.org/10.1038/nature24621.

Wylie, Alison (2020) Radiocarbon Dating in Archaeology: Triangulation and Traceability. In: Sabina Leonelli and Niccolò Tempini (eds) Data Journeys in the Sciences. Springer, Cham, pp 285–301. https://doi.org/10.1007/978-3-030-37177-7_15.