JKU

JOHANNES KEPLER
UNIVERSITY LINZ

JOHANNES KEPLER
UNIVERSITÄT LINZ

How can highly contextual data such as metabarcoding data relate to the world?

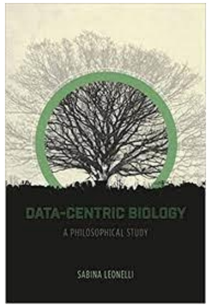
Aline Potiron
19 July 2021

JOHANNES KEPLER
UNIVERSITY LINZ
Altenberger Straße 69
4040 Linz, Austria
jku.at

Aims

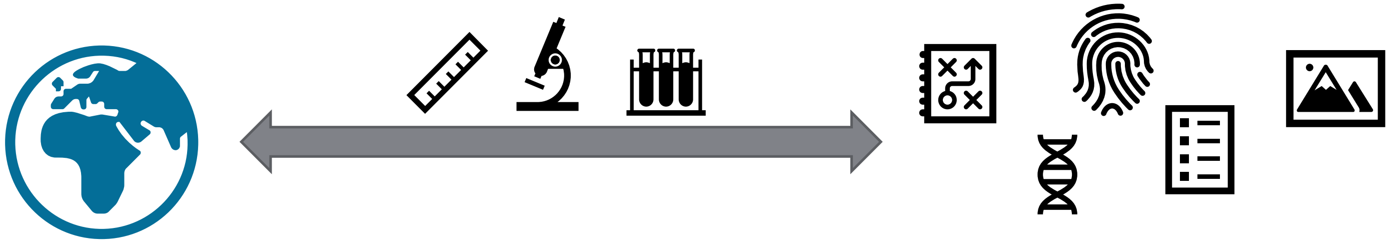
- I used a method widely used in microbiology and microbial ecology, **metabarcoding** (MB), to extend the understanding of the **relational view of data** developed by Sabina Leonelli in *Data-centric Biology: A Philosophical Study* (2016).
    - Leads to a refinement of this view of data by identifying a **particular role for representation** and sheds lights on the **investigative roles of different objects produced during the investigation.**

- The travel of MB data between different situations of scientific inquiry initiates an **epistemic shift from organism-centred biology to a more environment-centred biology.**
    - **Biological ontologies** capture the context of data production and constrain the representative power of data and thus their interpretational scope.



Leonelli (2016) *Data-centric Biology: A Philosophical Study.* The University of Chicago Press.

Introduction



Representational view Intrinsic properties	Objects	Relational view Role within the scientific inquiry
Are representations of a part of the world. They capture some properties of a phenomenon.	**Data**	Are “any product of research activities, ranging from artifacts […] to symbols […] that is collected, stored, and disseminated in order to be used as evidence for knowledge claims”.
Is a part of the world that has at least some mind-independent properties.	**Phenomenon**	Is the target of the scientific inquiry.
Are representations of more features of the phenomenon or the whole phenomenon.	**Models**	Are a way of ordering data so that they will represent a targeted phenomenon.
Is a more abstract conceptualisation of the world.	**Theory**	Is a source of knowledge that helps interpretation and is enriched by data and models.

Leonelli (2016) *Data-centric Biology: A Philosophical Study.* The University of Chicago Press.

Leonelli, S. What distinguishes data from models?. *Euro Jnl Phil Sci* **9,** 22 (2019). https://doi.org/10.1007/s13194-018-0246-0

Introduction

- The **relational view of data** addresses some of the problems of the **representational view**:
  - Distinction between data and models
  - Recognition that data are human-made, theory-laden but the representational view underestimates that in scientific inquiry, **producing data for themselves can be more important than the knowledge claim to be tested** *i.e.* than the representational content of the data.
    - e.g. Big Data science

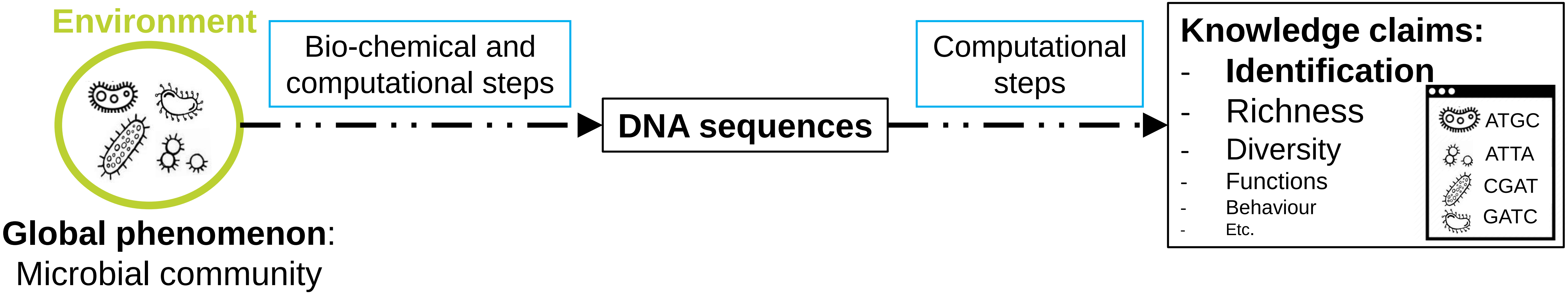
  ⇒ Here, data production seems to be the most important point. Data are supposed to serve as evidence for several claims sometimes about different phenomena.

  ⇒ **What do these data represent then?**

Leonelli, S. What distinguishes data from models?. *Euro Jnl Phil Sci* **9,** 22 (2019). https://doi.org/10.1007/s13194-018-0246-0

Introduction

- **Metabarcoding** (MB) is one of those methods where data production is important:



Environment

Bio-chemical and computational steps

DNA sequences

Computational steps

Knowledge claims:

- **Identification**
- Richness
- Diversity
- Functions
- Behaviour
- Etc.

Global phenomenon: Microbial community

- By generating **millions of DNA sequences,** MB is expected to “reveal” the underlying characteristics of the microbial community in general.
- Thus, MB can be considered as a **data-centred** method.

Introduction

- The **relational view of data** is best suited to the study of data-centric practices and thus to the study of **MB data.**
- Two features are **necessary** in this view for an(y) object to be data. An object must have:
  - (1) the **potential** to serve as **evidence** for sustaining knowledge claims and
  - (2) the **potential to travel** between different situations of scientific inquiry

NB: Situationism is a kind of **contextualism** BUT a situation gathers only those elements (whatever they are, events, objects, etc.) of the context that are **RELEVANT for the agent's current inquiry.**

Leonelli (2016) *Data-centric Biology: A Philosophical Study.* The University of Chicago Press.

Designed by Freepik

Problems

- Two questions I want to assess here:

1. If DNA sequences are produced for themselves, when and how do they acquire the potential to travel and to serve as evidence *i.e.* **when and how do these objects become data?**

The "meta" in MB stands for several organisms studied ***per each environment, date and/or conditions****. These parameters are thus considered relevant to explain the features of the microbial community and can be* ***considered as a part of the situation of scientific inquiry****.*

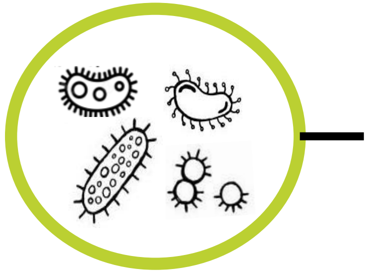
⇒ The retrieved DNA sequences are thus **highly contextual**

2. In this condition, **how can MB data be used as evidence in different situations of scientific inquiry?**

- **Aim:** What does this analysis bring to the relational view of data?

What constitutes data in metabarcoding?

Environment



Global phenomenon: Microbial community

Sequential processes	Results – potential data
Sampling	Material samples *e.g.* water
DNA extraction	Total DNA samples
DNA barcode amplification	Material DNA barcodes
DNA sequencing	Computerized DNA sequences
DNA sequences processing	Curated DNA sequences
DNA sequences modelling	Models

Knowledge claims

- All of these objects have the **potential to travel**: they can be stored locally or in virtual/physical databases and they can be sent by post or emailed to researchers for use in different inquiries.
- If we assume that there is continuity between these objects, after all, the material samples contain the DNA that will ultimately feed into the models that will ground conclusions about the microbial community, then they all have the **potential to serve as evidence**.

⇒ To speak in terms of potential amounts to consider **all objects produced in MB as data**.

What constitutes data in metabarcoding?

- I propose a finer distinction of the role of data within scientific inquiry: data as objects that **realize the travel** and data as objects that **actually serve as evidence**.

Data in the relational view developed by Leonelli

			Potential to travel	
			Travel Realized	
			Yes	No
Potential to serve as evidence	**Evidence role realized**	Yes		Curated DNA sequences Models
		No	Computerized DNA sequences	Material samples Total DNA samples Material DNA barcodes

⇒ It seems that **objects lose their ability to travel when they realize their potential to serve as evidence.**

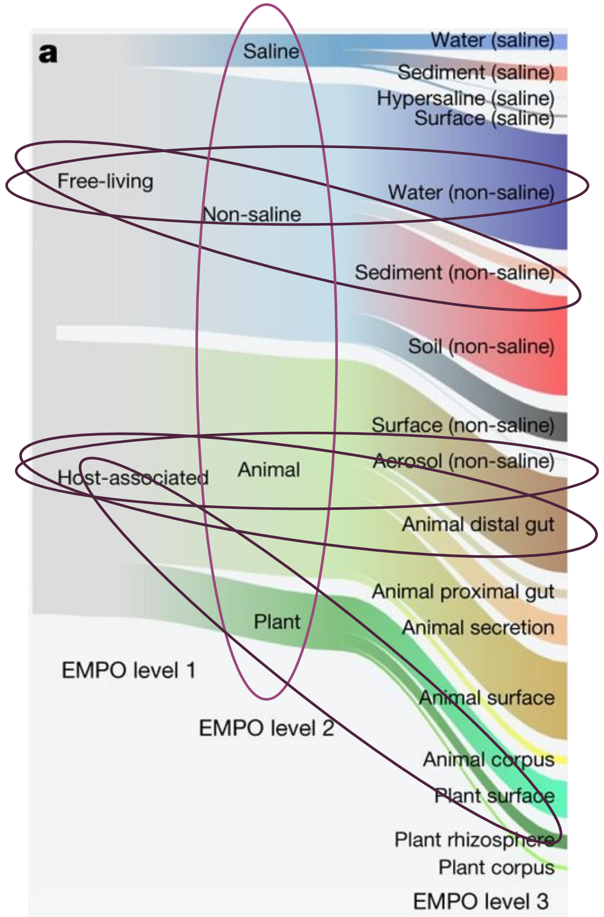
What constitutes data in metabarcoding?

- What do these objects gain when they realize their potential to serve as evidence?
- One possible answer: they **represent a targeted phenomenon**
  - The representational power of material samples and non-curated DNA sequences is only in the background of the scientific activity. These objects are handled in order **to secure their reliability to serve as evidence**.
  - However, in order to produce curated sequences and models, **scientists make choices.**
    - Those choices have an impact on **the shape that the data will take to serve as evidence**.
    - They will also have an impact on the **more precise part of the phenomenon that data can represent** by giving them a **biological significance**.

- The relational view of data does not deny a representative role of a given data, it simply **denies that what that data represent cannot change between situations of scientific inquiry**.

Leonelli, S. What distinguishes data from models?. *Euro Jnl Phil Sci* **9,** 22 (2019). https://doi.org/10.1007/s13194-018-0246-0

How can metabarcoding data serve as evidence in different situations of scientific inquiry?



- E.g.: **The Earth Microbiome Project** (EMP)
  - Collaborative project between researchers from around the world to understand microbial communities across environments globally.
  - Gathered 27,751 samples from 97 independent studies for a total of 2.2 billion (non-curated) DNA sequences.
  - Independent studies: Context of data production and interpretation.
  - 2017 paper: Contexts of data production and context of data interpretation.

  ⇒ **The context of interpretation changes but not the context(s) of data production** *i.e.* non-curated DNA sequences do not change but curated DNA sequences and models drawn from them change between the 97 studies and the 2017 paper.

Thompson, L., Sanders, J., McDonald, D. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551,** 457–463 (2017). https://doi.org/10.1038/nature24621

The Earth Microbiome Project logo was designed by Eamonn Maguire of Antarctic Design, Fair use, https://en.wikipedia.org/w/index.php?curid=35013837

How can metabarcoding data serve as evidence in different situations of scientific inquiry?

- Constitute **different situations of scientific inquiry** between the independent studies and the 2017 paper:
  - Different interpretative contexts.
  - Different research questions: “data are gathered primarily in service of separate questions rather than a single theme”.

- How does these **highly contextual data** travel across **different situations of scientific inquiries**?
  - **Standardised** material operations increase the **reliability** of data.
  - Non-curated DNA sequences do not have a fixed representational content – or have a wide spectrum of potential representation.
  - A common **ontology.**

Thompson, L., Sanders, J., McDonald, D. *et al.* A communal catalogue reveals Earth’s multiscale microbial diversity. *Nature* **551,** 457–463 (2017). https://doi.org/10.1038/nature24621

How can metabarcoding data serve as evidence in different situations of scientific inquiry?

- Ontologies capture what scientists consider **relevant in the context of data production for their scientific explanation.**
  - E.g.: pH, Temperature of the environment have an impact on the microbial community.
- By settling these parameters, ontologies **limit the type or aspect of the phenomenon that the data can represent** *i.e.* they restrain the representative power of the data.
- By linking and hierarchising these context parameters, ontologies also **constrain the range of possible interpretations of the data** *i.e.* the number of situations in which they can travel.
- Traditionally, biology accumulates knowledge about organisms. Either individually or in population.
- In MB, **the unit of investigation is the environment**. Data and knowledge are accumulated in relation to environments. Conclusions drawn from specific environments can be used to draw conclusions about related environment(s).

Conclusions

- MB is widely used in microbiology and microbial ecology. At the beginning of the investigation process, **the production of data takes precedence over their biological meaning**, it is a **data-centred method**. Hence I used the **relational view of data** to analyse the MB data.
- However, for the MB data, the decoupling of (1) the potential to travel and serve as evidence and (2) the realisation of these potentialities shed lights on the **investigative roles of the different objects produced during the investigation.**
- The representational power of MB data is constrained by the **ontologies** developed to capture certain parameters of the context judged by scientists to have a role in the biological explanation of a phenomenon.
- MB and other ‘omics’ techniques thus produce **an epistemic shift in the consideration of the unit of analysis from organisms to their environment,** reflected in these ontologies.

Thank you for your attention