Formats & Protocols in (Open) Science

Nathanael Sheehan

Fotis Tsiroukis

University of Exeter

Technical University of Munich

Philosophy of Open Science for diverse research environments

European Research Council
Established by the European Commission

What are Data Formats?

- Data formats are the encoding structure used to store and record information (Kitchin, 2025).
  - Not to be mistaken with *data types* (e.g. numerical, date, string) or *data structures* (e.g. array, lists)
- In a relational view of data, data “*are formatted and handled in ways that enable its circulation among individuals or groups for the purpose of analysis*” (Leonelli, 2016, p.78)
  - One key aim of Open Science has been to develop “open” data formats to challenge against “proprietary” data formats.

.csv

.GeoTIFF

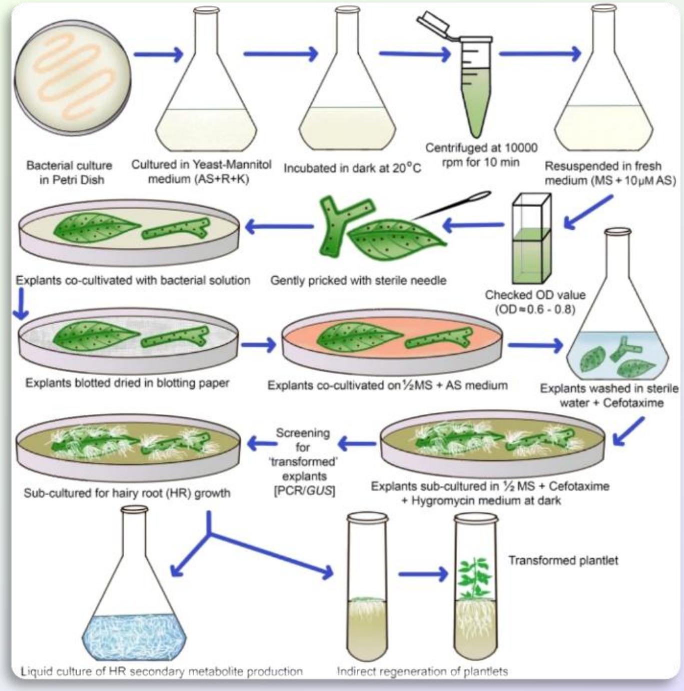
.FASTA

.GeoJson

What are Scientific Protocols?

- *Operational Perspective:* Protocols scaffold coordination of complex collaborative work.
- *Scientific Protocols*: Not only operational, but epistemic
  - Means of: *replicability, validating results, communicating know-how* (Lynch, 2001)
  - **Domain-dependent**

Biology			Agronomy		
Molecular	Genomics	Biochemistry	Plant Science		Conservation
PCR	Kits & Pipelines	Mixtures & Buffers	Propagation Protocols	Cultivation Protocols	Accession Protocols


Bacterial culture in Petri Dish
Cultured in Yeast-Mannitol medium (AS+R+K)
Incubated in dark at 20°C
Centrifuged at 10000 rpm for 10 min
Resuspended in fresh medium (MS + 10μM AS)
Checked OD value (OD≈0.6 - 0.8)
Gently pricked with sterile needle
Explants co-cultivated with bacterial solution
Explants blotted dried in blotting paper
Explants co-cultivated on ½MS + AS medium
Explants washed in sterile water + Cefotaxime
Explants sub-cultured in ½ MS + Cefotaxime + Hygromycin medium at dark
Screening for 'transformed' explants [PCR/*GUS*]
Sub-cultured for hairy root (HR) growth
Liquid culture of HR secondary metabolite production
Indirect regeneration of plantlets
Transformed plantlet

General Rule

Codified information about coordinated technical action (in a specific domain)

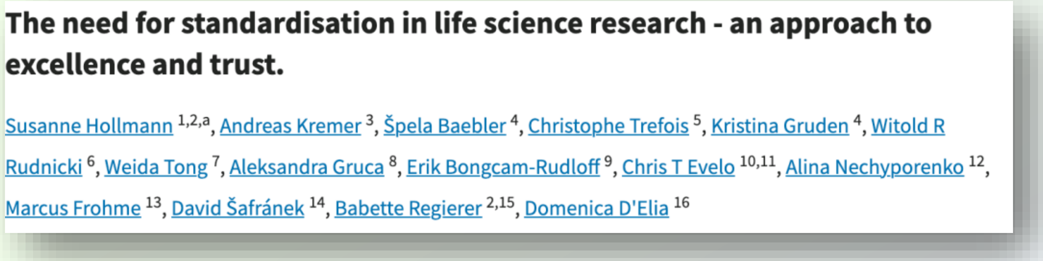
- ○ Clear
- ○ Step-wise
- ○ Sequential

The “Doxa”

Data Formats and Protocols are a means of achieving **interoperability** across *diverse research environments* and *communities of practice*

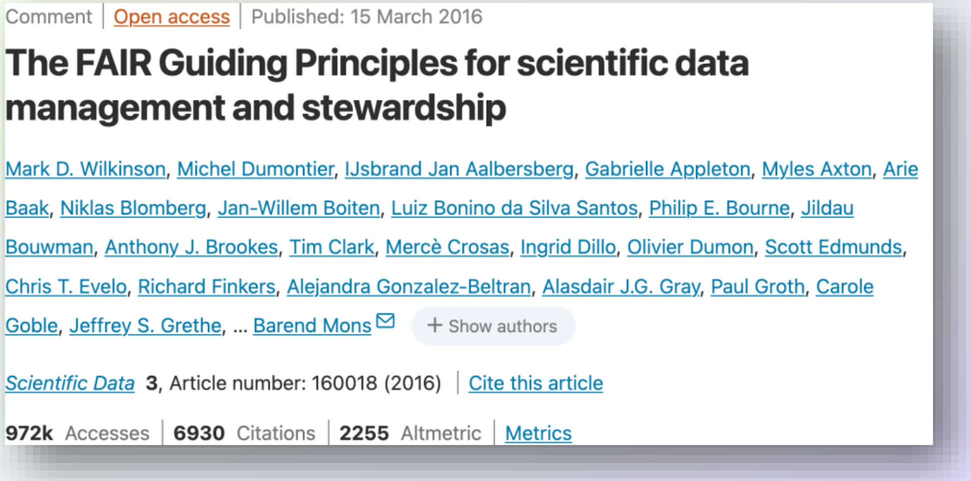
Interoperability has become a contemporary scientific *virtue* (particularly in the life sciences), with the expectation of structuring and linking information from diverse sources.

It is perceived as a source of hope for bringing ***stability, replicability and reproducibility*** to scientific practices.



The need for standardisation in life science research - an approach to excellence and trust.

Susanne Hollmann[1,2,a], Andreas Kremer[3], Špela Baebler[4], Christophe Trefois[5], Kristina Gruden[4], Witold R Rudnicki[6], Weida Tong[7], Aleksandra Gruca[8], Erik Bongcam-Rudloff[9], Chris T Evelo[10,11], Alina Nechyporenko[12], Marcus Frohme[13], David Šafránek[14], Babette Regierer[2,15], Domenica D'Elia[16]



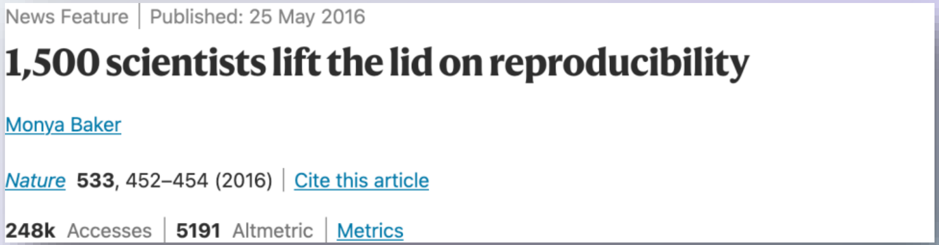
Comment | Open access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, ... Barend Mons ✉ + Show authors

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

972k Accesses | **6930** Citations | **2255** Altmetric | Metrics



News Feature | Published: 25 May 2016

1,500 scientists lift the lid on reproducibility

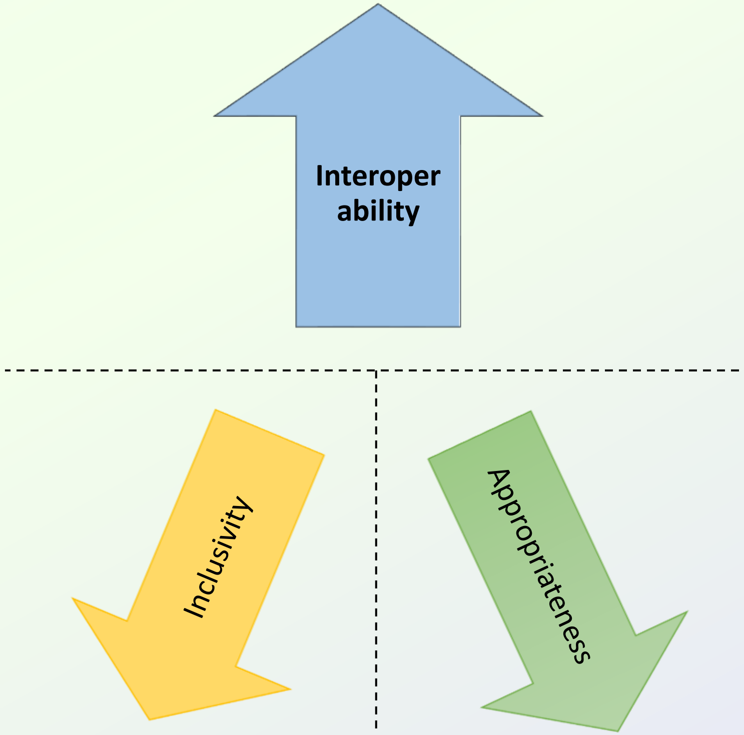
Monya Baker

Nature **533**, 452–454 (2016) | Cite this article

248k Accesses | **5191** Altmetric | Metrics

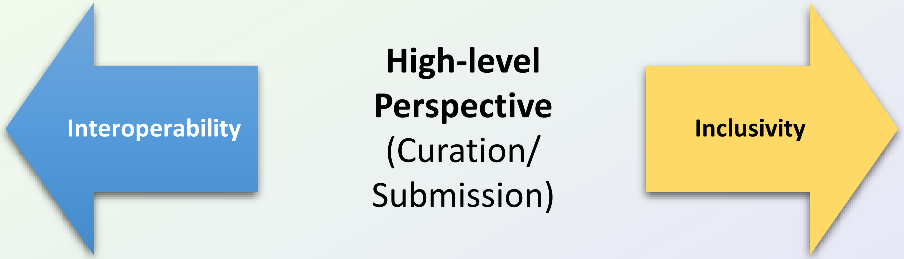
Our Subversion

However, we will argue, this creates a *Trade-off* between **Interoperability** and **Inclusivity/Appropriateness**



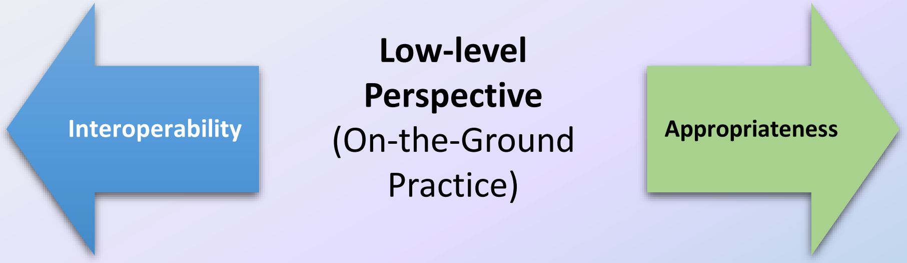
Case_Study_1: *Critical*

A Case of Data Format Exclusion



Case_Study_2: *Exemplary*

Translating between Global and Local Protocols



Studying the Trade-off

We believe a useful approach to study the tensions in the trade-off is by focusing on three simple questions centered on **exclusion:**

***Who* is excluded?**

Humans (professional roles, expertises, skills, cultures, identities)

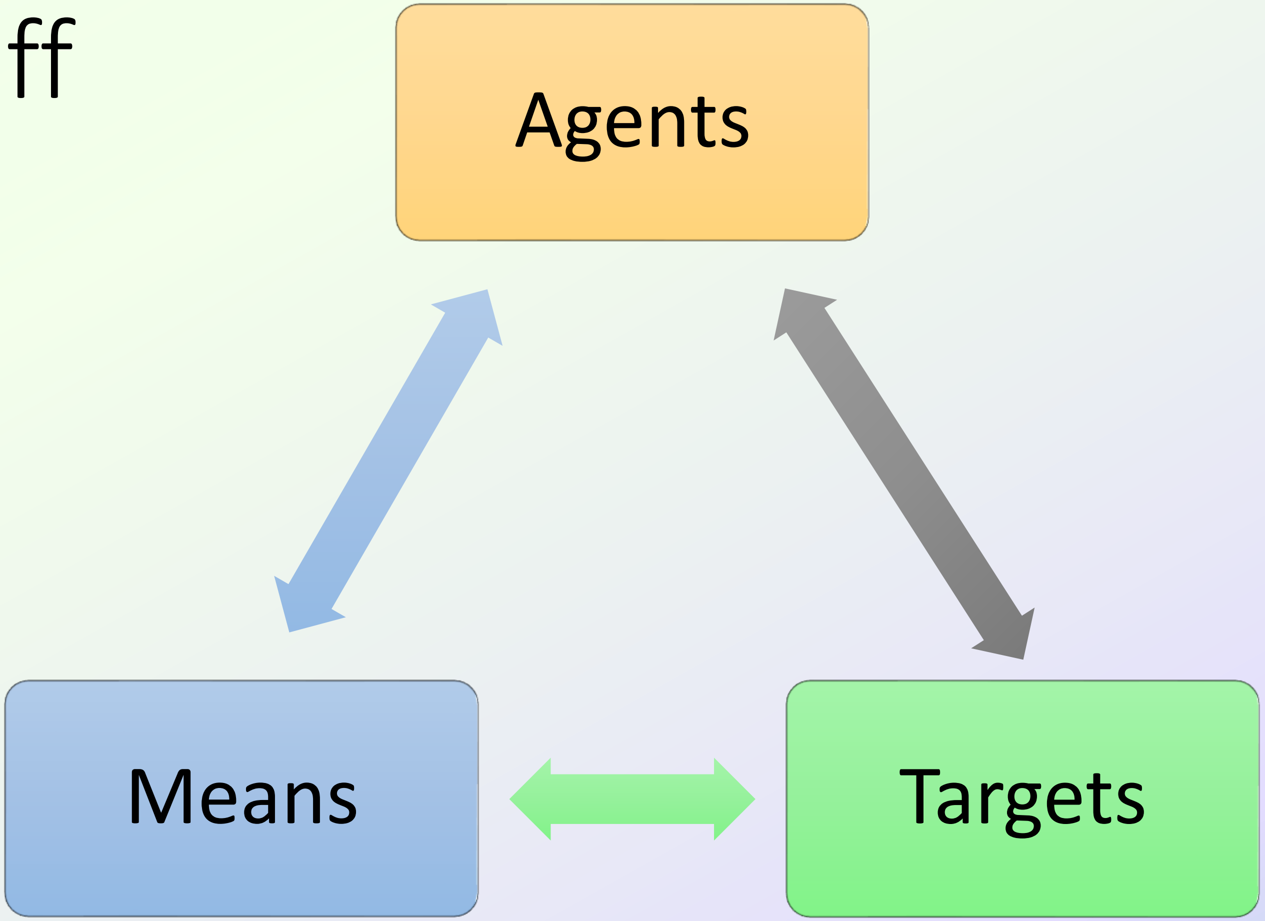
Non-Human life forms (species, ecosystems, conditions)

***What* is excluded?**

Actions, behaviors and processes

***Where* are exclusions?**

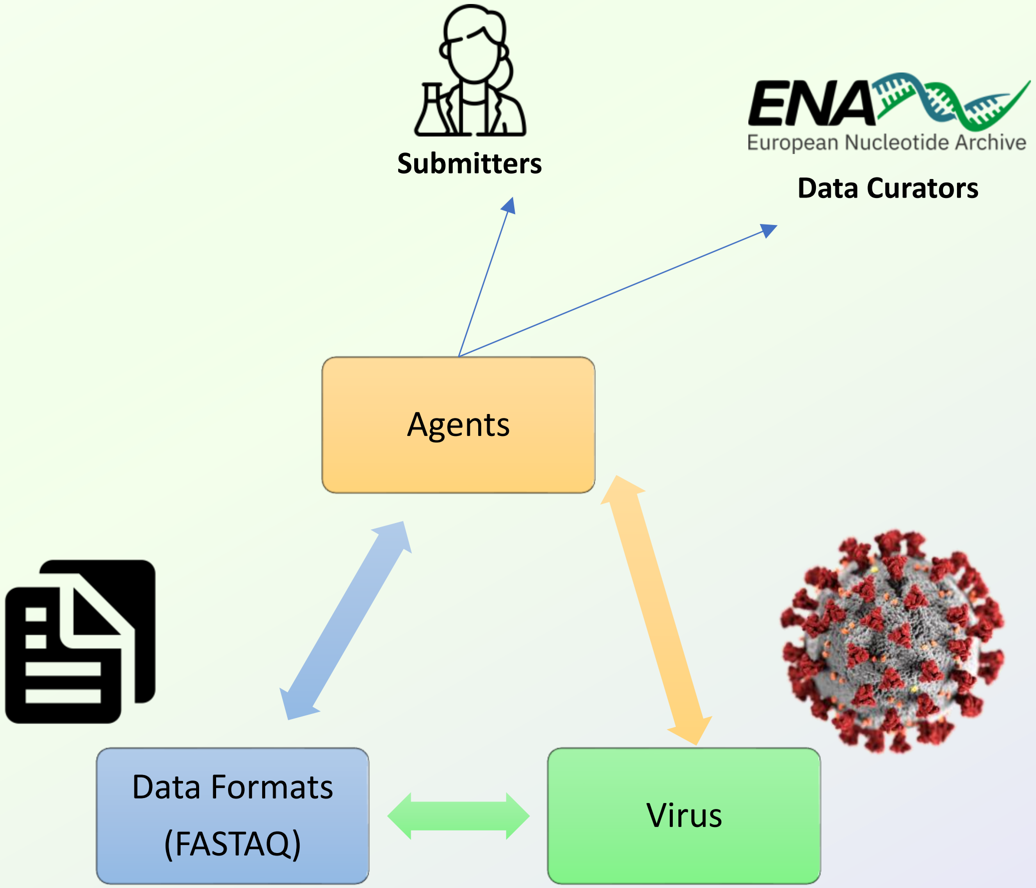
Where are the sites in which the exclusions are formed and where do the exclusions travel to affect who and what is excluded?



- Phenomena & Targets of Research
- Methods & Media of Research

+ Actors & Subjects of Research

Case Study 1: *Data Formats* at The European Nucleotide Archive



Submitters

ENA European Nucleotide Archive

Data Curators

Agents

Data Formats (FASTAQ)

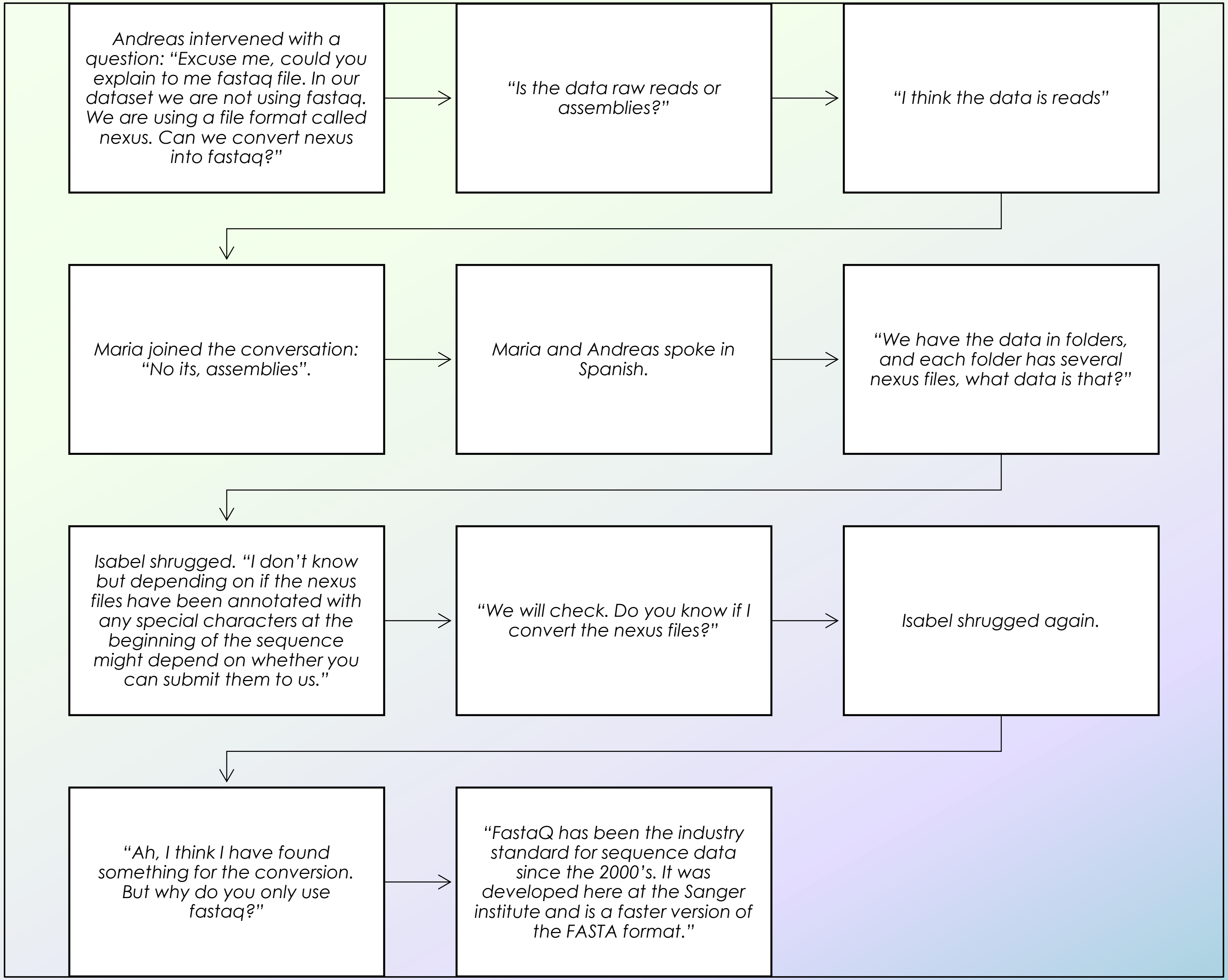
Virus

Context:

- *The European Nucleotide Archive*
- *The European Bioinformatics Institute*
- *The Wellcome Genome Campus*

Format Exclusion in action

“People think it's a really easy thing to just send files from there to there. But no, because we have **different types of files, different types of processes**.
Data Curator 3, ENA.



Andreas intervened with a question: “Excuse me, could you explain to me fastaq file. In our dataset we are not using fastaq. We are using a file format called nexus. Can we convert nexus into fastaq?”

“Is the data raw reads or assemblies?”

“I think the data is reads”

Maria joined the conversation: “No its, assemblies”.

Maria and Andreas spoke in Spanish.

“We have the data in folders, and each folder has several nexus files, what data is that?”

Isabel shrugged. “I don’t know but depending on if the nexus files have been annotated with any special characters at the beginning of the sequence might depend on whether you can submit them to us.”

“We will check. Do you know if I convert the nexus files?”

Isabel shrugged again.

“Ah, I think I have found something for the conversion. But why do you only use fastaq?”

“FastaQ has been the industry standard for sequence data since the 2000’s. It was developed here at the Sanger institute and is a faster version of the FASTA format.”

Participatory Information Format exclusion

Participatory Informational Format Exclusion (PIFE) refers to the exclusion of epistemic agents from participating in scientific systems - such as databases or repositories- due to their data being in a different format to what the systems accept.

Participatory Information-Format Exclusion occurs when:

(i) Agent **A** seeks to contribute information/data to a scientific database or infrastructure (thus exercising participatory epistemic agency),

(ii) The information/data held by **A** does not conform to the format standards required by the infrastructure, and

(iii) These format standards are governed by *informational asymmetries* (**S1** and **S2**), wherein the institutions controlling the standards possess *surplus* informational resources and interpretative control that **A** does not have.

(iv) Leading to **A** not participating in knowledge production.

Implications of PIFE

PIFE reveals inequalities in global data sharing

- E.g. During the COVID-19 pandemic countries with access to cutting-edge sequencing technologies, contributed most genomic data to global repositories.

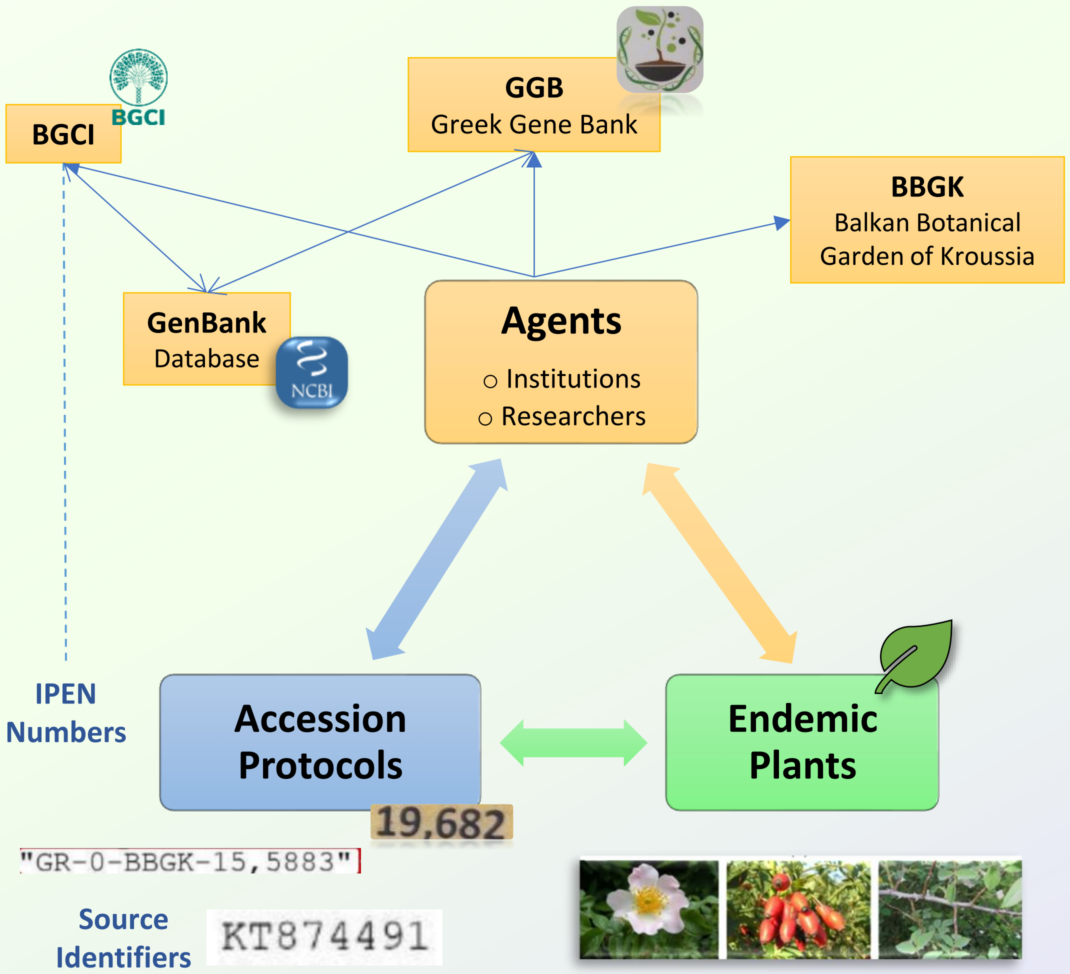
PIFE highlights Institutional Power

- PIFE shows how certain formats created in well-resourced environments dictate what counts as legitimate scientific knowledge.

PIFE occurs across various disciplines

- researchers in low-resource environments encounter significant barriers when engaging with dominant formats of data production and sharing.
- E.g. NETCDF for climate data banks like the Artic Data Archive.

Case Study 2: Plant Accession Codes @ IPBGR



BGCI

GGB
Greek Gene Bank

BBGK
Balkan Botanical Garden of Kroussia

GenBank
Database

Agents
- ○ Institutions
- ○ Researchers

IPEN Numbers

Accession Protocols

Endemic Plants

19.682

"GR-0-BBGK-15,5883"

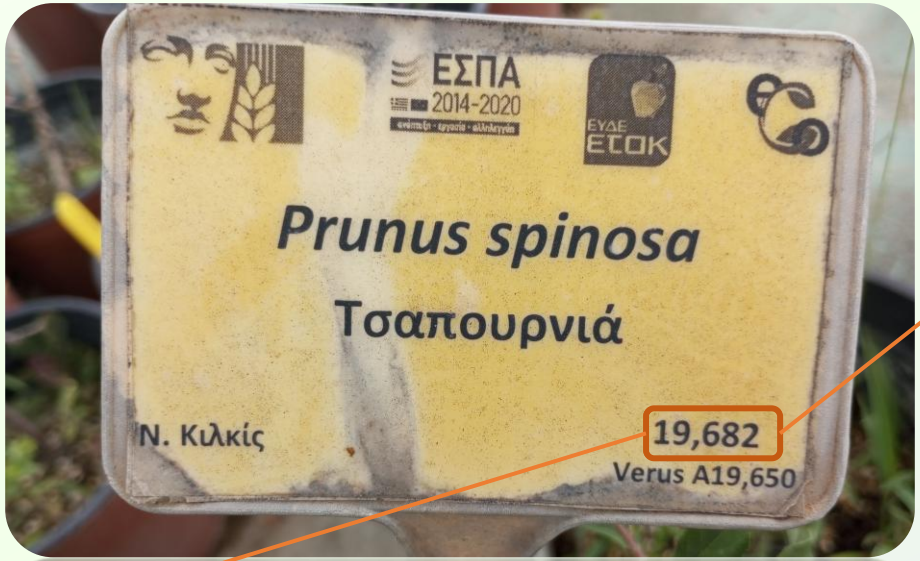
Source Identifiers

KT874491

Context:

- *IPBGR Thermi, Greece*
- *Endemic Species In Situ Collection (Outdoors & Greenhouse*
- *Greek Gene Bank Living Collections & Database*
- *Part of Systematic Botany & Genetic Assesment Research*

Local Identification Protocol



At point of collection

"if it has a name, it has a code"

Code

YY,
000(00)
(N)
(X)

Protocol

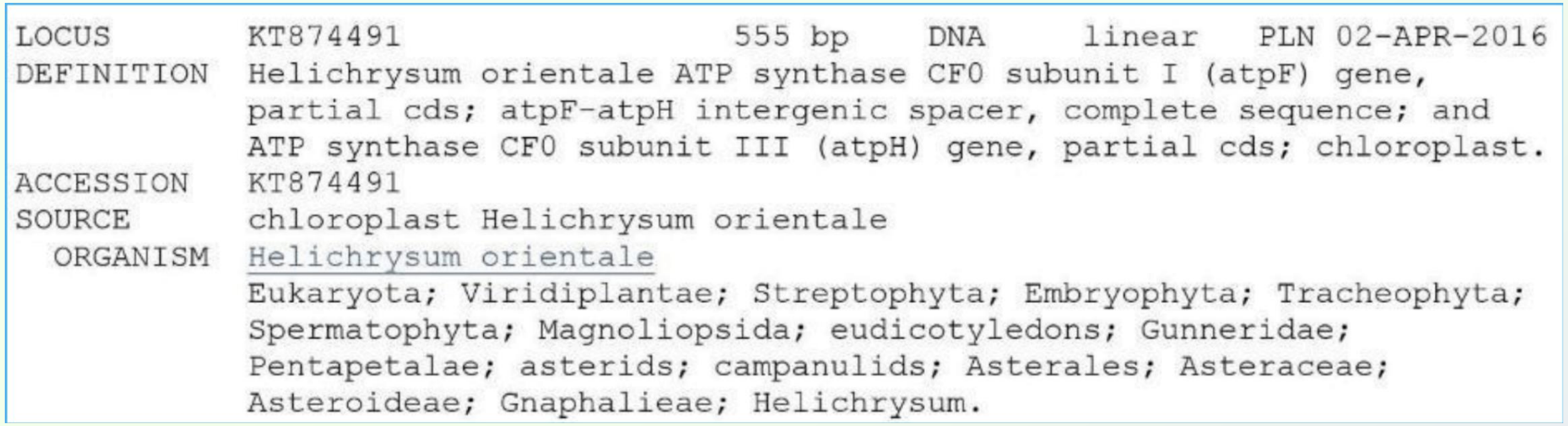
- Year of Collection (19,…)
- Serial Code (…,682) ascending order
- Generation Number (…-2) ascending order
- Sample Letter: indicating sample (…-A)

Properties:

- Used in all practices involved in the institutes.
- Adaptable/Expandable (but not really replaceable or malleable).
- Connects local practices
- Requires manual work (non-automated, reliance on few people)

Plant Accession Numbers

Genebank Standards
for Plant Genetic Resources
for Food and Agriculture



```
LOCUS       KT874491                 555 bp    DNA     linear   PLN 02-APR-2016
DEFINITION  Helichrysum orientale ATP synthase CF0 subunit I (atpF) gene,
            partial cds; atpF-atpH intergenic spacer, complete sequence; and
            ATP synthase CF0 subunit III (atpH) gene, partial cds; chloroplast.
ACCESSION   KT874491
SOURCE      chloroplast Helichrysum orientale
  ORGANISM  Helichrysum orientale
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae;
            Pentapetalae; asterids; campanulids; Asterales; Asteraceae;
            Asteroideae; Gnaphalieae; Helichrysum.
```

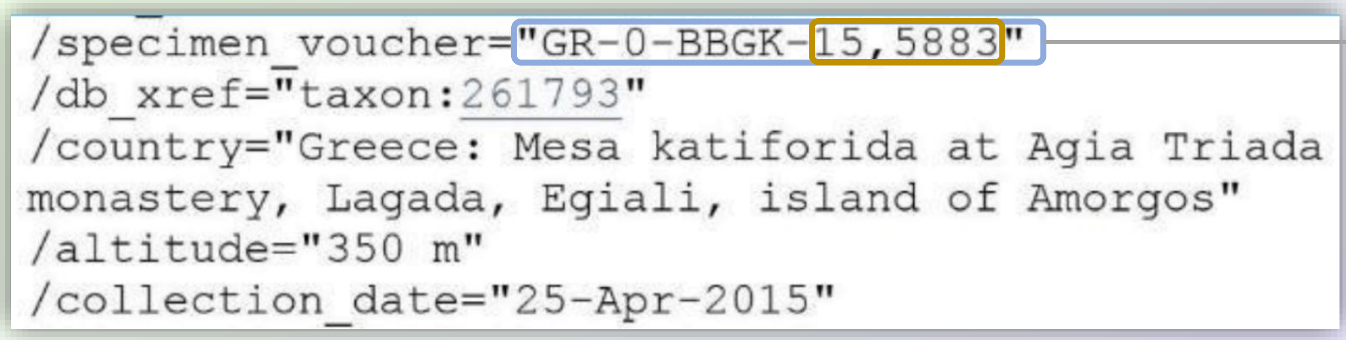
NCBI GenBank Database

Source Modifiers Table File

Requires Geolocation:

A contentious issue
(disrupts local ecosystems)

Not *exclusion*, but *inclusion of something* ***inappropriate*** to the local situations



```
/specimen_voucher="GR-0-BBGK-15,5883"
/db_xref="taxon:261793"
/country="Greece: Mesa katiforida at Agia Triada
monastery, Lagada, Egiali, island of Amorgos"
/altitude="350 m"
/collection_date="25-Apr-2015"
```

Local to Global Plugin:

-Country abbreviation
-Restriction Code
-Institute Abbreviation
-Local AN number [YY,000-(x)]

➔XX-n-XXXXNN,NNNN-n(or X)
➔e.g. GR1BBGK19,654-2

The local code protocol is embedded inside the metadata of the submission document

Generalizing from the Case

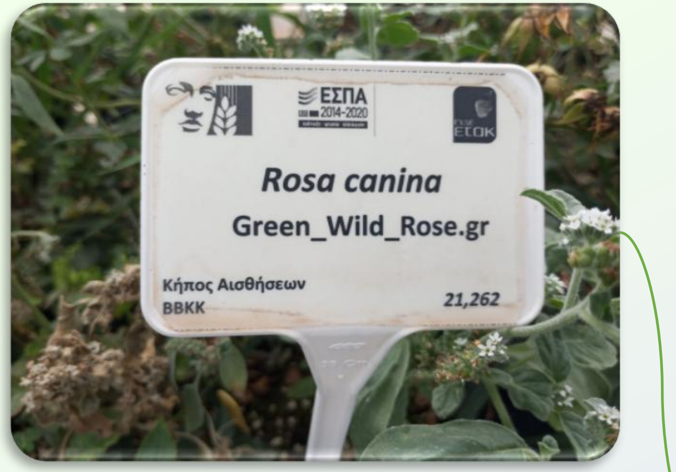
Mediation:
The local identification protocol mediates between **global standards** (*IPEN*) and the **practices** (*collection, characterization, propagation, cultivation*) around the **target plants** (*Rosa canina, Prunus etc.*).

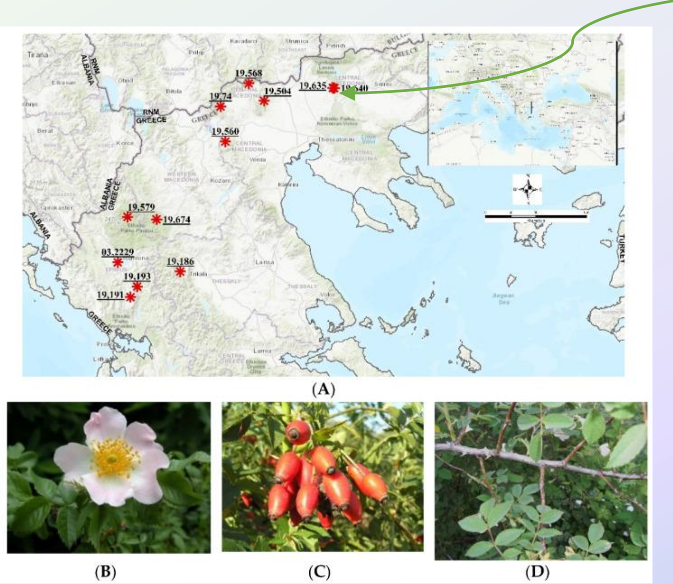
***"Protocollage"**:*

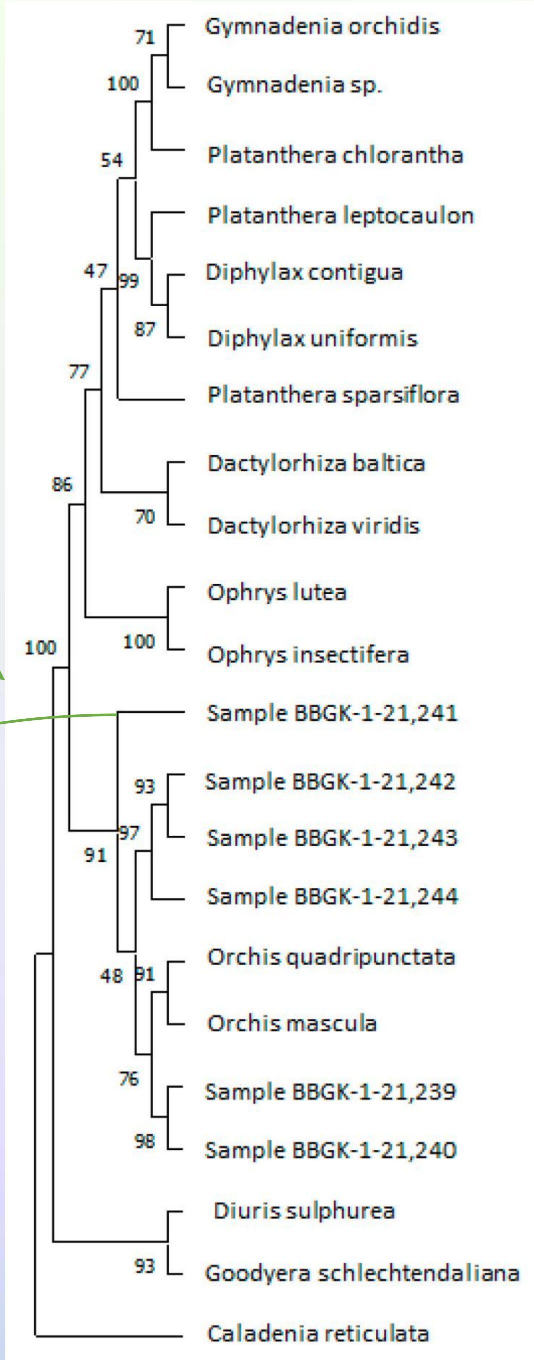
- Protocols usually come in assemblages, forming **systems, procedures, pipelines** (botanical collection, genotypic analysis, data storage).
    - Indentification protocols "contains" the whole pipeline (they are used during start and end)
- Protocols are **modular** and can *plug-in* to other protocols or combine with others to form **composites** (Local AN -> IPEN)

Epistemic Importance:

- Id protocols fix objects in order to track them.
- Protocols codify and formalize research practices on target plants. They represent action/behaviors (process ontology)
- Analogy to models: *If models represent causal relations in target systems, protocols represent actions.*







Open Questions

- **Data Formats**: How can we tackle the challenges of PIFE in ways that goes beyond just making data formats technically interoperable?
    - How should responsibilities be distributed between technical infrastructure developers, who often dictate data formats based on interoperability standards, and researchers operating in low-resourced environments?
- **Research Protocols**: How can we become "protocol aware"?
    - What properties protocols capture? What practices do they leave aside? What expert knowledge is implied?
    - When do global standards create conflicts with local practice?

Common Issues:

- Over-Reliance on Standards
- Reification (Dupre & Leonelli, 2022)