Proof & Provenance

A biobank data model for coordinating and sharing phages and results

Jan Zheng, Ben Temperton, Jon Iredell

A growing number of databases have been created for the phage world, including PhagesDB, the Viral Host Range Database, the International Phage Therapy Database, and several "catalogue" databases like DSMZ, ATCC, and NCTC. They're amazing projects, but their phage data aren't generated from lab data. Ideally we'd want lab data to directly inform us:

- **What do we know about a phage**, is it **what we think it is**, and **is it pure**?
- **Where did a phage come from**? Where did it go? And what's been done with it?

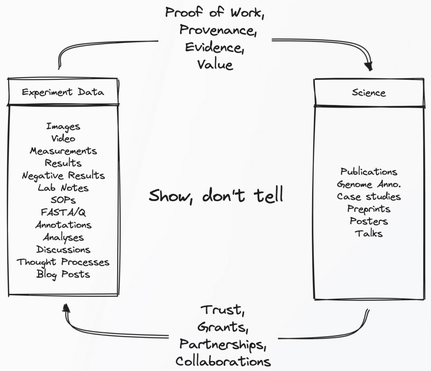

Fig 1. Show, don't tell.

Data generated from experiments can be used to provide proof of work and provenance for published science.

For example, the plaque assay data, genomics analyses, and TEM images created for a phage would be combined to paint a full picture of what we know about that phage.

Inspired by @jackbutcher's "Sell your Sawdust" thread.

When we send, receive, and use our phages, we want to trust that a phage is and does what it is. As the field moves towards GMP-grade phages, all lab processes need to be documented through well-defined Standard Operating Procedures (SOPs). SOPs should define strict lab data collection requirements, which can then be used to derive the characteristics of our phages. This data would establish a baseline of characterization for any phages that are shipped between labs. When other labs work with our phages, they should be able to add their data to our phage, and vice versa.

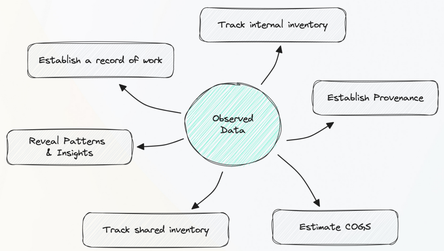

Fig 2. Observed data collected from experiments helps us understand our phages.

We can use plaque assay, isolation, and accession data to tell us where a phage is, what it does, and where it's been sent. The data can even help us estimate how much effort, materials, and lab time it took to produce a phage.

The more data we're able to collectively gather about our phages, the more we'll know about the characteristics of our phages.

We should aim to create a catalogue, or "Pokédex," of phages. The catalogue should provide enough information for another lab to make their independent assessments about the quality or usefulness of a phage, before they receive it.

By sharing our SOP documentation, wet lab data and bioinformatics data, and aggregating them into phage catalogues, we'll be one step closer to establishing "Material Safety Data Sheets" for phages.

Proof of Phage

What do we know about the phage?

Drogon

Name:	Drogon
Isolation Strain:	atcc-39327-paer
ID:	cpl-0404-paer
Sequencing ID:	cpl-seq-1534
Sequencing URL:	/cpl.bio/cpl-seq-1534.json
Family:	Podovirus
Genome size:	38,123
GC %:	46%

Safety Notes
...

Production Notes
...

Event log
... Assay / experiment data added here

Fig 3. An illustration of what a "Pokédex-like" phage entry might look like.

All characteristics about a phage, like host range, should be directly calculated from its SOP-defined assay results.

By deriving all characteristics from openly available lab results, phages should be much easier to audit.

Provenance: Proof of History

Where did it come from → Where did it go?

Phage Passport

Name:	Drogon
Owner:	Citizen Phage Library
Discovered by:	Daenerys #cpl523
Isolated by:	Citizen Phage Library
Isolated date:	May 5, 2023
Isolated location:	/// index.again.snake
Isolation source	Wallaby
Isolation Host:	P. aeruginosa
ID:	cpl-0404-paer

Events
jul-8-23: #evt-1422 CPL -> Phdus
oct-12-23: #evt-1562 CPL -> IPATH
nov-15-23: #evt-1774 sequencing data added

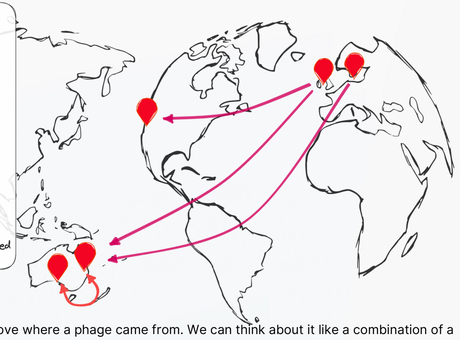

Fig 4. "Provenance" is our ability to prove where a phage came from. We can think about it like a combination of a Birth Certificate, a CV, and Passport of everything it's ever done, and everywhere it's ever been sent. As each lab receives, characterizes, and uses the phage, they'll add to the phage's CV by contributing data to it.

By openly sharing and collaborating on lab results, labs can show that they've discovered, characterized, and contributed to certain phages. By adding to the characteristics of a phage, labs can prove the usefulness and value of various phages in a biobank. When labs use a phage in experiments or use them to treat patients, the case studies and published data will accrue with the phage, contributing to the phage's "CV". When multiple lab share phages and contribute data, we can establish a "map" of where a phage was sent, and what was done to it, thus establishing provenance for phages.

Distributed provenance w/ data sharing & replication

Data is validated with Merkle tree hashing or proof-of-work mechanisms

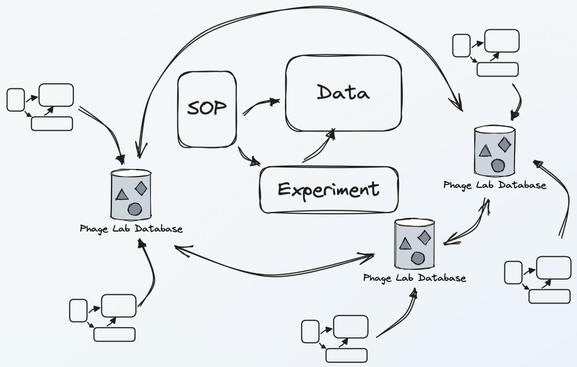

Fig 5. All experiment data should be generated according to one or more well-defined SOPs.

This data should be stored in a distributed data system (as opposed to a single database), so labs can fully control what they store and what they share.

Labs could run their own replicated nodes within a network of "federated" databases, similar to Mastodon.

In the future, cryptographic ideas like "Merkle tree hashing" or "ZK rollups" could support anonymous yet independently-verifiable data and lab results.

Our data is only as good as the lab work that generates the data. This is why using shared protocols and strict data requirements is so important. With documented protocols, we'll know which groups "report temperature as Celsius, and which as Fahrenheit," and which groups use which processes and techniques. As methods improve, we'll be able to deprecate older results, know which datasets have switched to newer methods, and which SOPs generated which data.

All lab data is added as raw files and lab notes. Rather than using a schema-centric form-based data entry system, we should allow for tagging files and notes in a way that can be automatically aggregated and processed down into searchable graph and vector databases. When phages undergo training, drift, or engineering, these variations and "drift" can be tracked and mapped using various tools like from their genomic distances in the database. We can use a combination of traditional and NLP/LLM-based data extraction techniques with a document database like CouchDB to extract information from raw text and images into the database.

This type of data model, rather than a traditional SQL or noSQL database gives us the flexibility to extract and create relationships and structured information from any lab and biobank-generated raw files and text-based data like photos, lab notes, or machine-generated data, when applied to a biobank data platform context.

Lab data workflow

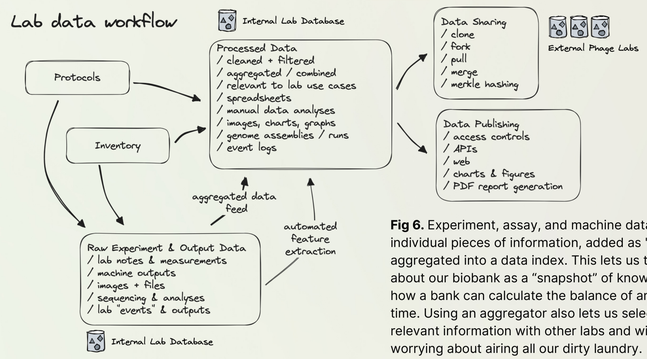

Fig 6. Experiment, assay, and machine data are collected as individual pieces of information, added as "data events", and then aggregated into a data index. This lets us treat anything we know about our biobank as a "snapshot" of knowledge in time, similar to how a bank can calculate the balance of an account at any point in time. Using an aggregator also lets us selectively filter and share only relevant information with other labs and with the public, without worrying about airing all our dirty laundry.

Biobank data model

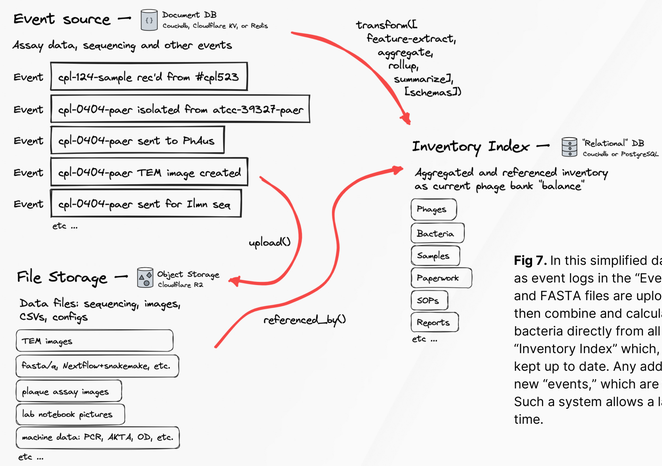

Fig 7. In this simplified data model, all lab results data is collected as event logs in the "Event source" database. All files, like CSVs and FASTA files are uploaded to the file storage database. We can then combine and calculate information about our phages and bacteria directly from all the lab results and uploaded files, into an "Inventory Index" which, like a bank account balance, is always kept up to date. Any additions, edits, or corrections are added as new "events," which are then recalculated into the Inventory Index. Such a system allows a lab's data to be auditable at any point in time.

Data points ideally tracked by biobanks

It's more than just assay data and genomics

Lab Inventory
Samples
Phages
Hosts
Batches
Boxes
Materials

Processes
Versioned Protocols
Methods
Lab Instructions & Wiki
— Where to find data
— who to ask
— what to do in case of X

Observations
Raw lab notes
Spontaneous thoughts
Photos and images
Sound, video, zoom recordings
CSVs, PDFs, Word, Excel docs
Machine outputs + sequencing files

Paperwork
Transfer agreements
Shipping logs, updates
Invoices
Nagoya paperwork
Legal templates + notes

Access & Usage
Access log
Aliquot / materiel consumption
Citations & Publications
Phage & Host Availability
Phage Production Notes

Fig 8. Biobanks will need to track more than just assay results and genomics outputs. Most LIMS and ELNs try to track a subset of lab inventory or notes, but are either too expensive or become too constrained. Both lab inventory and lab notes need to be treated as two sides of the same coin, as they both contribute to the "current state of a lab and its experiments." To support phage exchanges, labs will also need to exchange any usage, handling, and production instructions for phages, as well as all the paperwork and transfer agreements that come with the territory.

Biobank data application stack

(a database is not enough)

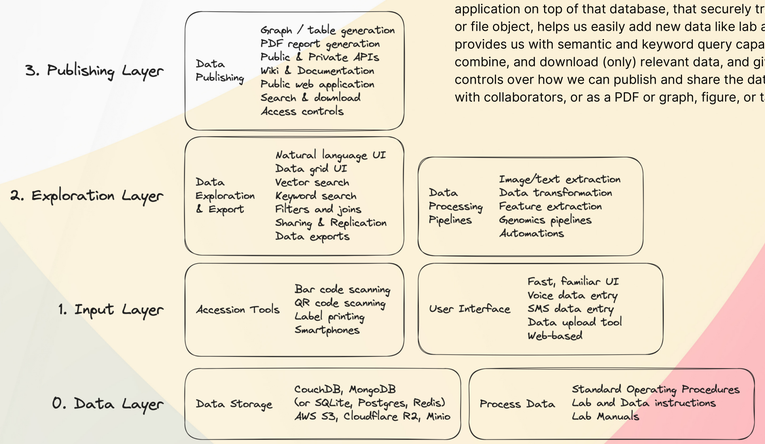

Fig 9. A biobank database is just the start. We also need a full-stack data application on top of that database, that securely tracks and uploads any data or file object, helps us easily add new data like lab and inventory notes, provides us with semantic and keyword query capabilities, lets us filter, combine, and download (only) relevant data, and gives us fine-grained controls over how we can publish and share the data — either with the web, with collaborators, or as a PDF or graph, figure, or table.

Based on Capsid & Tail **Issue 123: Proof & Provenance.** https://phage.directory/capsid/proof-provenance.
Thanks to Ben Temprton, Jon Iredell, and the rest of the Phage Australia team. Also many thanks to all those who helped over the years. All graphics created in Canva and Excalidraw.

Phage Directory · Phage Australia · The Westmead Institute · Citizen Phage Library