Data Management

Dr Courtney Lane
Lead Epidemiologist, Centre for Pathogen Genomics &
WHO Collaborating Centre for AMR

Dr Kristy Horan
AusTrakka Bioinformatician

SeqAsia Thailand, Online tutorials
2025-06-10

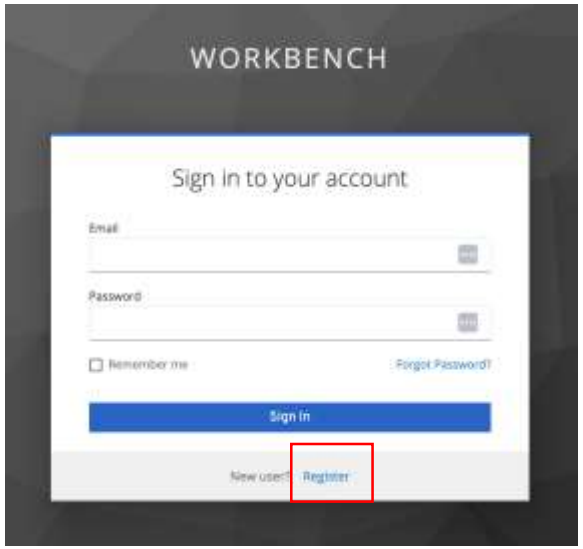
Preparing for Practical Sessions: Server access

BEFORE the next session (17th June)

Go to https://workbench.nbt.or.th

Set up an account

- Verify your e-mail address



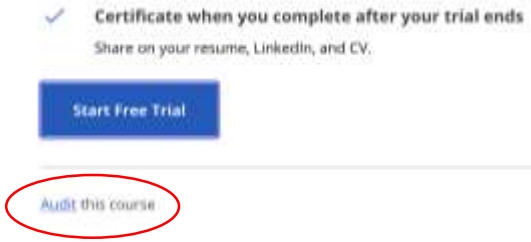
Wait for approval

- Try to login in 48 hours (you may not get an email)
- If you can't login after 48 hours, or any other issues e-mail worawich.pho@biotec.or.th

Preparing for practical sessions: Unix/Bash

BEFORE the first practical exercise session (24th June)

If you are not already familiar, please complete an online course on Unix / Bash



Options:

- https://www.w3schools.com/bash/index.php
- https://www.coursera.org/specializations/unix-and-bash-for-beginners
  - Click “try for free”, then “Audit this course”
- https://www.melbournebioinformatics.org.au/tutorials/tutorials/unix/unix/

Please complete the exercises, not just reading the material.

You have two options:

- If you have your JupyterLab access set up, open “Terminal” (recommended for at least some)
- Directly in the interactive panels

Data Management

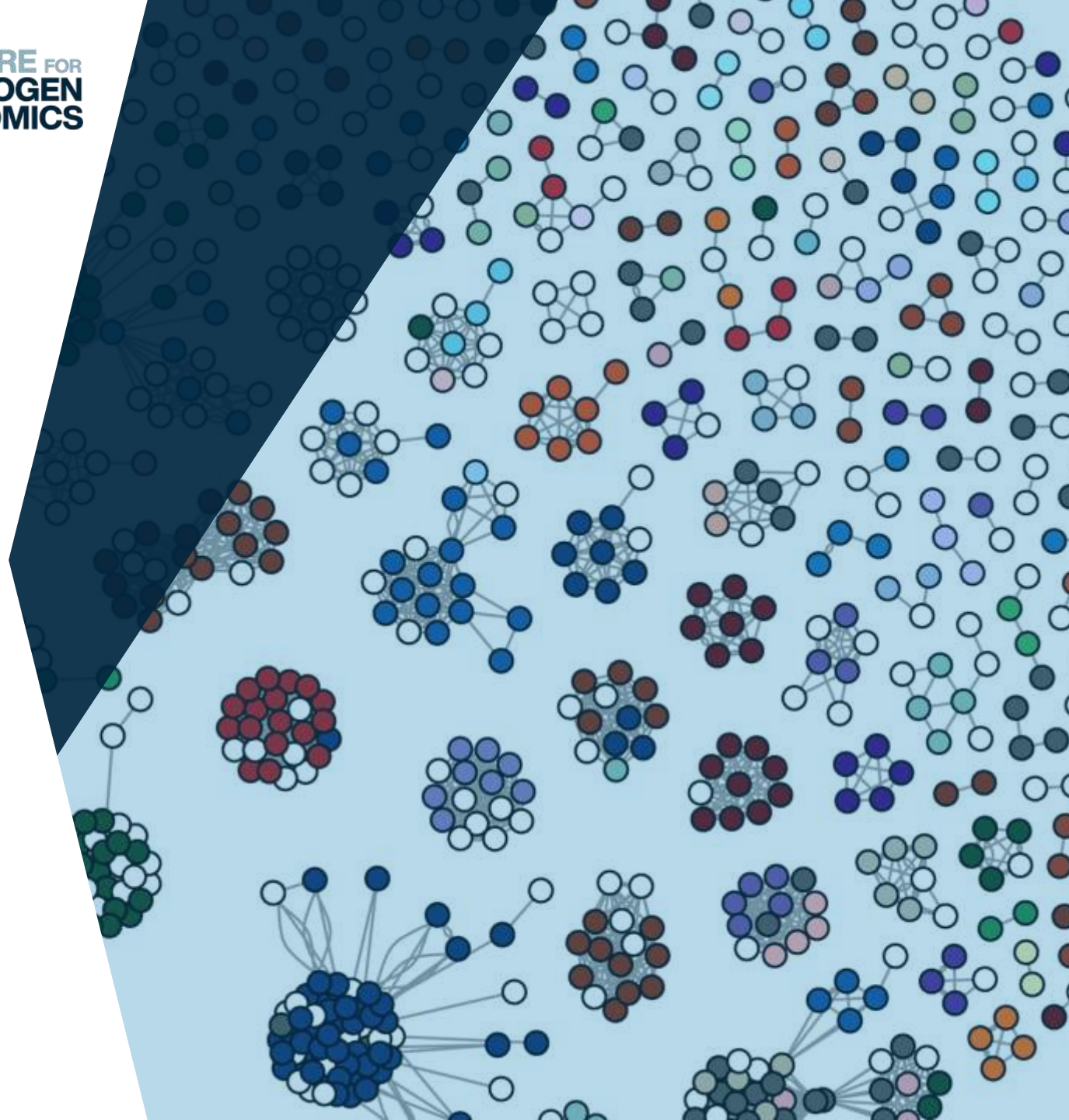
Content

Kristy:

- What is data?
- Computational infrastructure
- Software & databases
- Data management pipelines

Courtney:

- Core Metadata
- Enhanced clinical & epidemiolocal data
- Practical tips for collection of epidemiological & clinical data



Coffee Break

Reminder – explanation of pre-work for next week at end

Core “Metadata”

What is “metadata”?

- “Data that provides information about other data”
- Core pieces of information without which it is difficult to understand your data
  - In this case sequence data

- Who was the sample collected from? When was it collected? How was it collected?

What core information would you want to know about your samples and sequences, and why?

Core sample & host metadata

Data field	Utility
Sample and host data	Core data that answers the “person, place & time” the sequence came from. Needs to be provided with the sample, or linked from another source.
Unique sequence identifier	Identifies a sequence across databases. Allows linkage to other data.
Organism / taxonomic identification	Fundamental for comparative genomics and for placing the sequence in the correct biological context
Geographical location of sampling	Specific location (country, state/province, town) the sample was originally collected. Vital for phylogeograpy, understanding spread of pathogens, and distribution of genetic diversity.
Date of collection	Date on which the sample was taken. Critical for temporal and evolutionary analyses, understanding genetic diversity, inferring relationships, and understanding genetic diversity over time.
Source of isolate Host species Specimen type	Describes the environment from which the organism was isolated (e.g. soil, wastewater), host organism (e.g. human, animal) and the specific sample type (e.g., blood, faeces, nasal swab). Helps in understanding the ecology of the organism, clinical presentations, and potential spread/sources of infection
Host identifiers (identifiable or de-identified)	Allows for the linking of multiple samples from the same individual, important for understanding intra—host diversity, longitudinal data.
Host demographics	Demographic data (e.g. for human, age and sex). Crucial for genomics epidemiological analyses to identify at-risk populations, understand disease spread and susceptibility.

Technical metadata

Data field	Utility
Sequencing and analysis methods	Additional information you will collate through the sequencing process, and will store for future reference. Provides understanding of how the data were generated and ensures the reproducibility and quality assessment of the genomic data
Sequence platform	Specifies the technology used to generate the sequence data (e.g., Illumina MiSeq, Oxford Nanopore MinION)
Library preparation method	Details of the protocol used to prepare the genetic material for sequencing (e.g., whole-genome shotgun, amplicon-based)
Assembly method	Describes the software and versions used to assemble the genome and perform other analyses (e.g., variant calling)
Sequence quality metrics	Includes measures such as sequencing depth (coverage), quality scores, and completeness
Public repository accession numbers	The accession numbers for the raw sequence data (e.g., in NCBI's Sequence Read Archive) and the assembled genome (e.g., in GenBank) ensure that the data is Findable, Accessible, Interoperable, and Reusable (FAIR)

Collection of core sample & host metadata

- Sample collection / referral forms
  - Received for all samples -> perfect for “core” data
  - Those collecting or referring **should** have access to all core data (but can be a challenge)

Annex 3.2 Example sample referral form with core surveillance data fields

This form would be completed by the diagnostic laboratory when referring isolates to the NRL for further testing, surveillance and biobanking. Additional epidemiological or clinical data may be requested on the isolate referral form or provided with electronically submitted routine surveillance data. See Section B2.3.3.

Form adapted from WHO Laboratory Quality Stepwise Implementation Tool *(54)*

Microbiological Isolate Referral Form – <*laboratory name*>

Laboratory use only
Sample ID: _____
Received date: _____

Referring laboratory details

Laboratory name: _____ Contact person: _____
Laboratory address: _____ Phone number: _____

Patient details*

Family name: _____ Given name: _____
National identifier: _____ Date of birth: _____ (dd-mm-yyyy) or age: _____
Sex: ☐ Male ☐ Female ☐ <other> Ethnicity~: ☐ <as required> ☐ Unknown ☐ Not stated
Usual residential address: _____ Province: _____
Phone number: _____
Patient location at sample collection: ☐ Outpatient ☐ Hospital inpatient ☐ Unknown
If outpatient ☐ General practice ☐ Other outpatient If inpatient: Hospital name: _____
☐ Hospital emergency ☐ Unknown ward/unit: ☐ ICU ☐ Other, specify: _____
☐ Residential aged care facility admission date: _____ (dd-mm-yyyy)

Specimen details

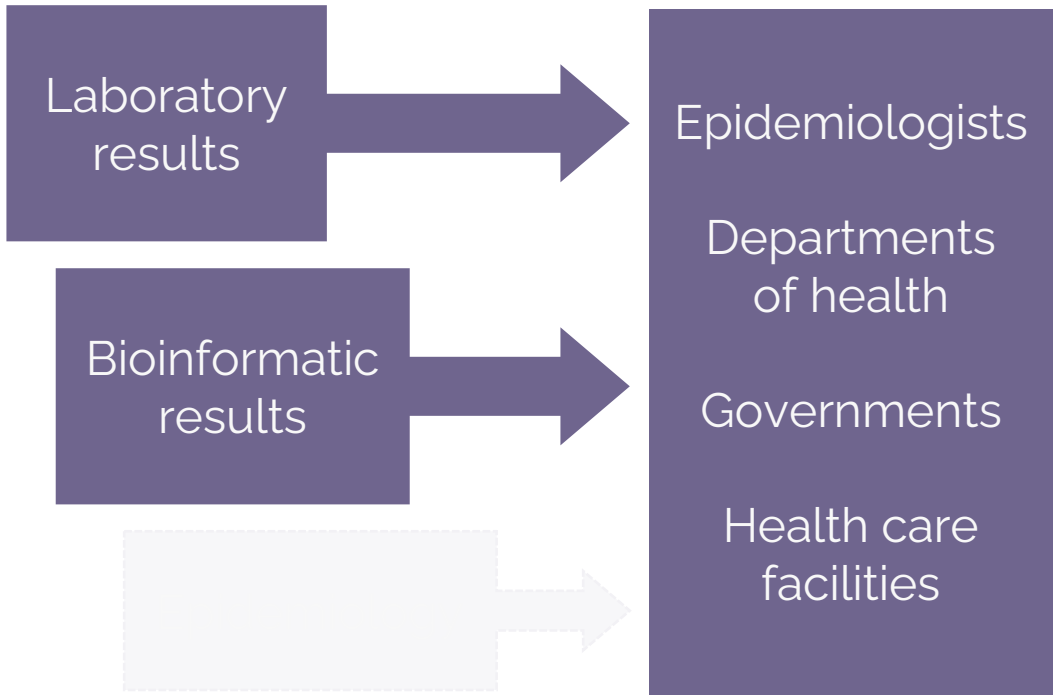
Collection date: _____ Collection date: _____ (dd-mm-yyyy)
Reason for collection: ☐ Clinically indicated Sample type: ☐ Blood ☐ Faeces ☐ Urine ☐ Sputum
☐ Screening ☐ Swab ☐ Pus ☐ Tissue ☐ Fluid
☐ Unknown ☐ Other, specify: _____
Sample body site: _____

Source: https://www.who.int/publications/i/item/9789290620365

Enhanced clinical and epidemiological data

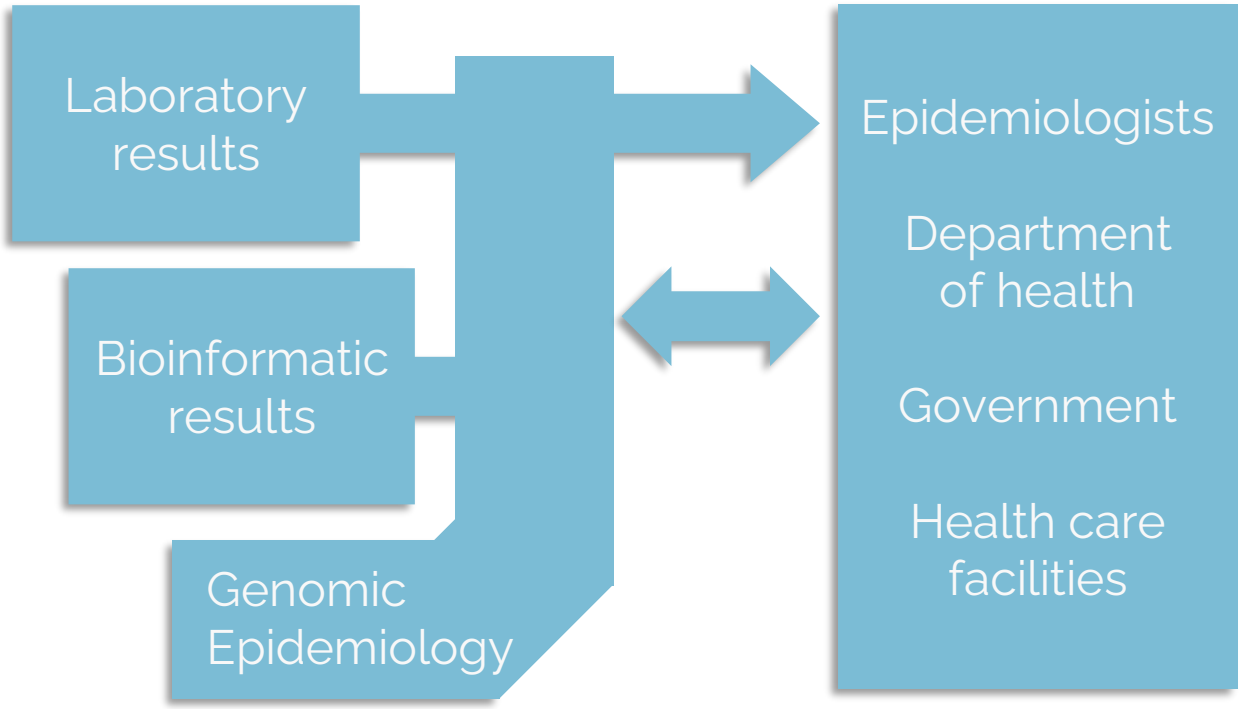
Genomic epidemiology in public health laboratories

Traditional



- Single sample static results
- One day data flow
- Combined analysis at end user level

Emerging structure



- Increased complexity of analysis
- Iterative analyses, 2-way communication and data flow with end-users
- Need for appropriate data storage

Enhanced clinical & epidemiological data

- Core data often not enough to answer specific public health or research questions

What kind of questions might you use genomic data for?

What additional data might you want collect to answer these?

(draw on own experience, or hopefully you have watched the first regional webinar)

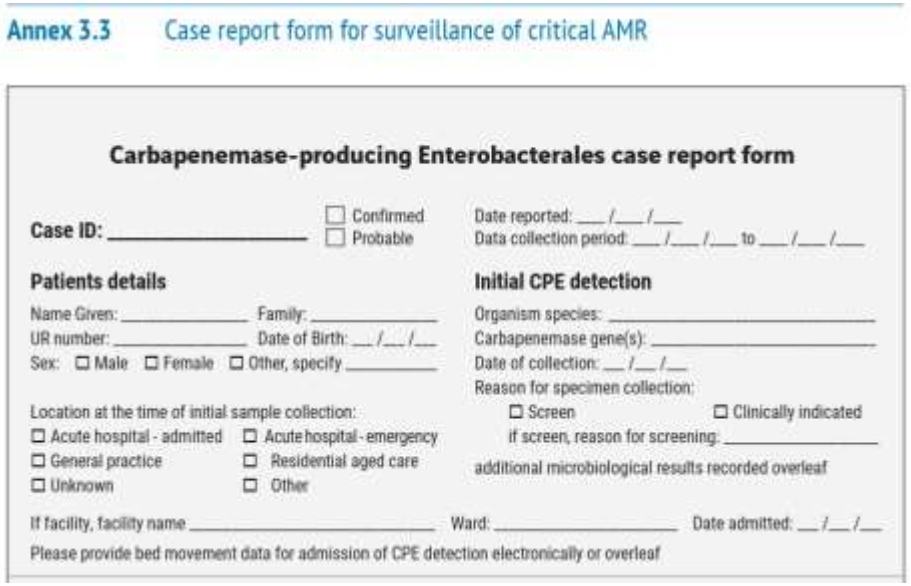
Examples of enhanced clinical & epidemiological data

Use case	Additional data to be collected, collated or stored
Monitor relationships between genetic mechanisms and phenotypic resistance	**Phenotypic antimicrobial susceptibility testing** results (Caution re: sample naming, is it really the same ”thing”?)
Understand disease importation risk e.g. determine frequency of importation and likely country of origins	**Patient travel history**; Origin of imported products
Understand transmission dynamics e.g. disease spread, at risk groups, and potential sources	Risk factor data relevant to pathogen & host e.g. travel, food or **hospitalisation history**; Livestock movements
Prospective outbreak detection	Outbreak identifiers, known epidemiological links
Identify strains of increased concern e.g. increased disease severity	**Clinical presentation/syndrome** (animal or human), relevant medical history
Monitor design of tests and vaccines e.g. identify potential vaccine escape mutants	Vaccination history (animal or human), other relevant medical history

Collection of enhanced clinical & epidemiological data

On sample collection / referral

- Who would have access to required data?
  - Infection control, surveillance officers,
  - Primary data collection (research studies, active surveillance)



Annex 3.3 Case report form for surveillance of critical AMR

Carbapenemase-producing Enterobacterales case report form

Case ID: _____ ☐ Confirmed ☐ Probable

Date reported: ___/___/___
Data collection period: ___/___/___ to ___/___/___

Patients details

Name Given: _____ Family: _____
UR number: _____ Date of Birth: ___/___/___
Sex: ☐ Male ☐ Female ☐ Other, specify _____

Location at the time of initial sample collection:
☐ Acute hospital - admitted ☐ Acute hospital - emergency
☐ General practice ☐ Residential aged care
☐ Unknown ☐ Other

Initial CPE detection

Organism species: _____
Carbapenemase gene(s): _____
Date of collection: ___/___/___
Reason for specimen collection:
☐ Screen ☐ Clinically indicated
if screen, reason for screening: _____
additional microbiological results recorded overleaf

If facility, facility name _____ Ward: _____ Date admitted: ___/___/___
Please provide bed movement data for admission of CPE detection electronically or overleaf

Source: https://www.who.int/publications/i/item/9789290620365

Linkage to other data sources

- Surveillance systems (notifiable diseases, sentinel surveillance)
- Hospital/clinic medical records
- Administrative data

- **Where else?**

- Note: Data about humans or companion animals is less more likely to involve linkage of of routine data like this, other sample types are more likely to use additional primary data collection

Data linkage

- What information is, or will be, stored across all databases?
  - Linkage can be very time consuming, if not automated

  - Multiple identifiers can reduce failure:
    - National / patient / study IDs
    - Name
    - Date of birth
  - As can probabilistic or “fuzzy” matching

BUT caution:

- Data types are often collected and stored at different levels
  - Does this host level information apply to all samples/sequences?
  - How do I identify data about the same sample or sequence, not just host?

Practical tips for data collection & storage

Best practices for epi/clinical data collection

Data collection should be:

- **Beneficial** to data collector and data provider
- **Standardised,** with variables having the same meaning
- **Repeatable** so that different data collectors would collect the same data if using the same method
- **Relevant,** with only necessary data collected, to answer specific questions
- **Timely** and as close to real-time as possible for timely decision-making and feedback
- **Disaggregated**, with data collected at the highest possible resolution

Practical advice for data collection

- **Standardised** (national) referral forms
- Use categorical fields where possible; **avoid free-text** fields
- Use a **standard ontology**, especially useful when sharing data to public databases (will have a session on this later)
    - Standard ontologies: https://genepio.org/ ; https://genepio.org/DataHarmonizer/template/pha4ge/SOP.pdf

- Provide a **data dictionary** to increase repeatability
- **Pilot** data collection tools / referral forms

Practical advice for data storage

- Metadata is stored separately to sequence data
  - Use laboratory information management system/databases where possible
    - Often not flexible enough to accommodate required data

  - Try to avoid using Excel to **store** data – very easy to introduce undetected errors
    - RedCap, EpiCollect, even Google forms are alternatives
    - BUT avoid storing re-identifiable, or sensitive data on cloud-based systems unless sure of data security and access

- Use validated rules for data entry

- Think about if and how you will keep data up to date

Key points

- Data is not just “Data”, there are different types of data
  - Raw sequences
  - Data about the host, sample & sequences
  - Derived data – analysis methods, results

- You need to be able to:
  - Know what data you need to collect and record
  - Link the different types of data to each other
  - Store and back-up different types of data

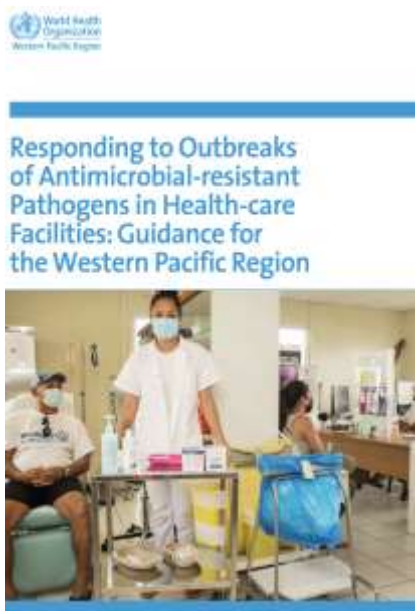
Questions?

Further reading

Example sample collection, referral and outbreak investigation forms:

https://www.who.int/publications/i/item/9789290620365

https://www.who.int/publications/i/item/9789290619970



Principles for pathogen genome data sharing:
https://www.who.int/publications/i/item/9789240061743

Genomic epidemiology ontology: https://genepio.org/

Next topic

17th June: QC of raw data (Theory)

24th June: QC of raw data (Exercise)

But there is some preparatory work…

"Davis, I'm beginning to think that quality control isn't your niche."

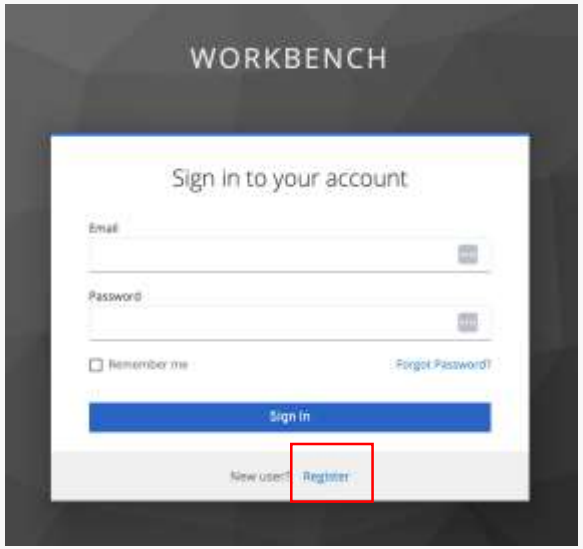
Preparing for Practical Sessions: Server access

BEFORE the next session (17th June)

Go to https://workbench.nbt.or.th

Set up an account

- Verify your e-mail address



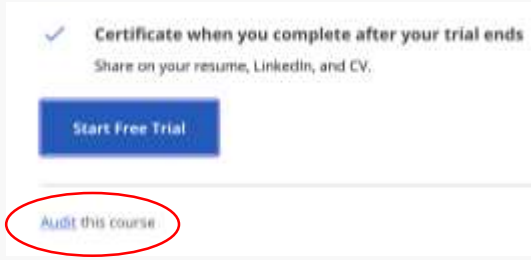
Wait for approval

- Try to login in 48 hours (you may not get an email)
- If you can't login after 48 hours, or any other issues e-mail worawich.pho@biotec.or.th

Preparing for practical sessions: Unix/Bash

BEFORE the first practical exercise session (24th June)

If you are not already familiar, please complete an online course on Unix / Bash



Options:

- https://www.w3schools.com/bash/index.php
- https://www.coursera.org/specializations/unix-and-bash-for-beginners
  - Click “try for free”, then “Audit this course”
- https://www.melbournebioinformatics.org.au/tutorials/tutorials/unix/unix/

Please complete the exercises, not just reading the material.

You have two options:

- If you have your JupyterLab access set up, open “Terminal” (recommended for at least some)
- Directly in the interactive panels