

# Metadata and submitting sequences

Geert van Geest  
Interfaculty Bioinformatics Unit, UniBe  
Training group, SIB

# Question

# Data life cycle



# Sequencing data life

## DANGER ZONE

- **No standardized metadata** - info gets lost
- **Limited backups** - technical failure leads to data loss

Data at sequencing center

Basic QC

Data short-term stored for analysis

Analysis + annotation

Data at repository (SRA)

Publication



# Get out of the danger zone

- Use your data management plan
- Plan data submission as part of the experiment
- Start the submission soon in the project

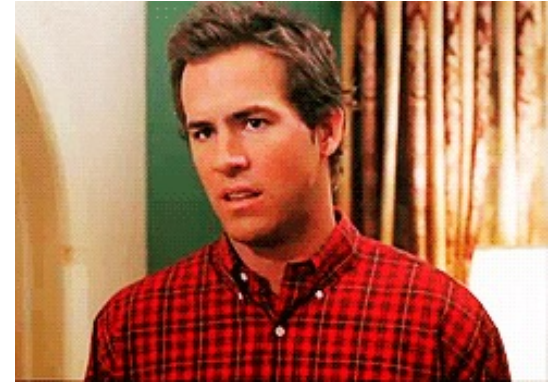


# How?

- Submit it to the sequence read archive of one of the INSDC databases:
  - Describe (metadata)
  - Upload

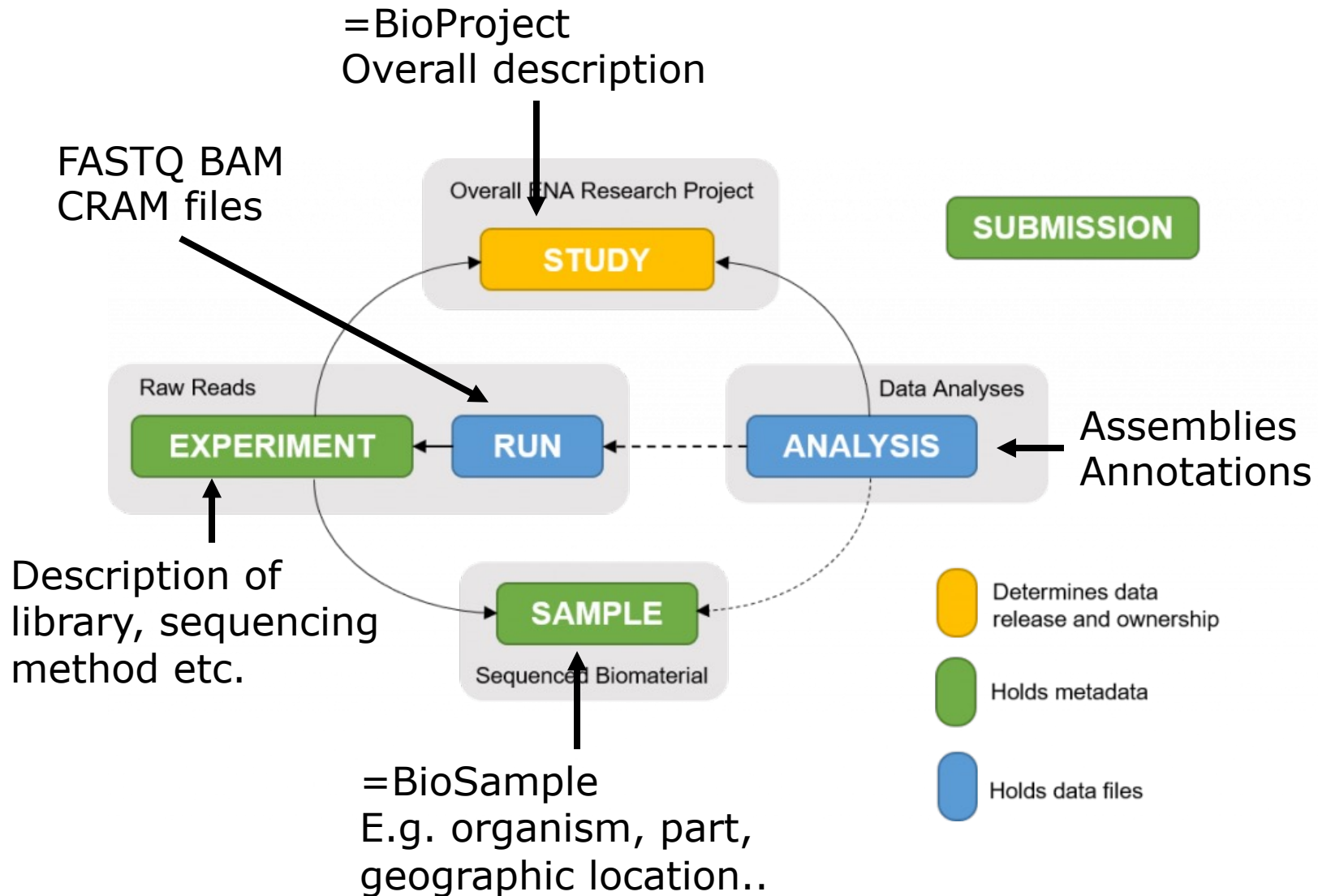


# Metadata



- Annoying:
  - **standardized** description of **unique** data, samples and experiments
  - Everything is already in the **paper!**
- But essential:
  - Enables finding
  - Enables re-use
- Data is useless without context!

# ENA database





# Transferring data

- Sequence data files are LARGE
- Possible protocols:
  - Webin File Uploader (GUI)
  - File Transfer Protocol (FTP)
    - Command line
    - FileZilla
  - Aspera



# md5 checksums

- Timeouts and failed transfers are common
- FASTQ files can become useless
- Solution: md5 checksum - changes if file changes

