EMBL-EBI

ENA European Nucleotide Archive

Automating a SARS-CoV-2 ENA submission tool with Nextflow

Zahra Waheed and Ahmad Zyoud

European Nucleotide Archive
European Molecular Biology Laboratory - European Bioinformatics Institute

Who are the European Nucleotide Archive (ENA)?

The ENA:

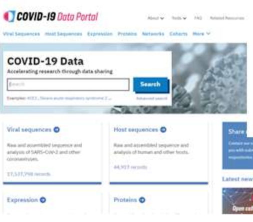
- Is a global, open-access nucleotide sequence repository.
- Covers non-sensitive raw sequence data, sequence assembly information and functional annotation for all organisms.
- Is the European arm of the International Nucleotide Sequence Database Collaboration (INSDC), which includes NCBI and DDBJ
- Mirrors data between NCBI and DDBJ

The ENA & SARS-CoV-2

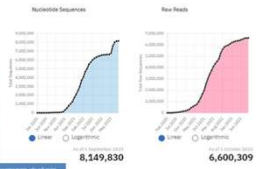
During the COVID-19 pandemic an unprecedented volume of SARS-CoV-2 data was submitted to the ENA. SARS-CoV-2 submissions now make up ~24% of all ENA raw sequence reads, which have been shared from 105 countries.

The COVID-19 Data Portal

- All public SARS-CoV-2 sequences and raw reads submitted to the ENA feed into the COVID-19 Data Portal
- This was launched by EMBL-EBI in April 2020 in response to the pandemic
- The Portal brings together a variety of SARS-CoV-2 related biological datatypes and visualisation tools for global access and analysis


COVID-19 Data Portal
About Tools FAQ Related Resources
Viral Sequences Host Sequences Expression Proteins Networks Cohorts More
COVID-19 Data
Accelerating research through data sharing
Search
Viral sequences
Host sequences
Expression
Proteins
Share
Latest news
BY-COVID

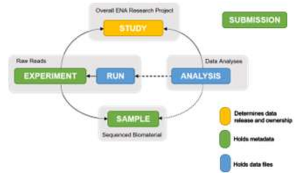
The COVID-19 Data Portal



Statistics for SARS-CoV-2 data submitted to INSDC

The SARS-CoV-2 Drag and Drop Submission Tool

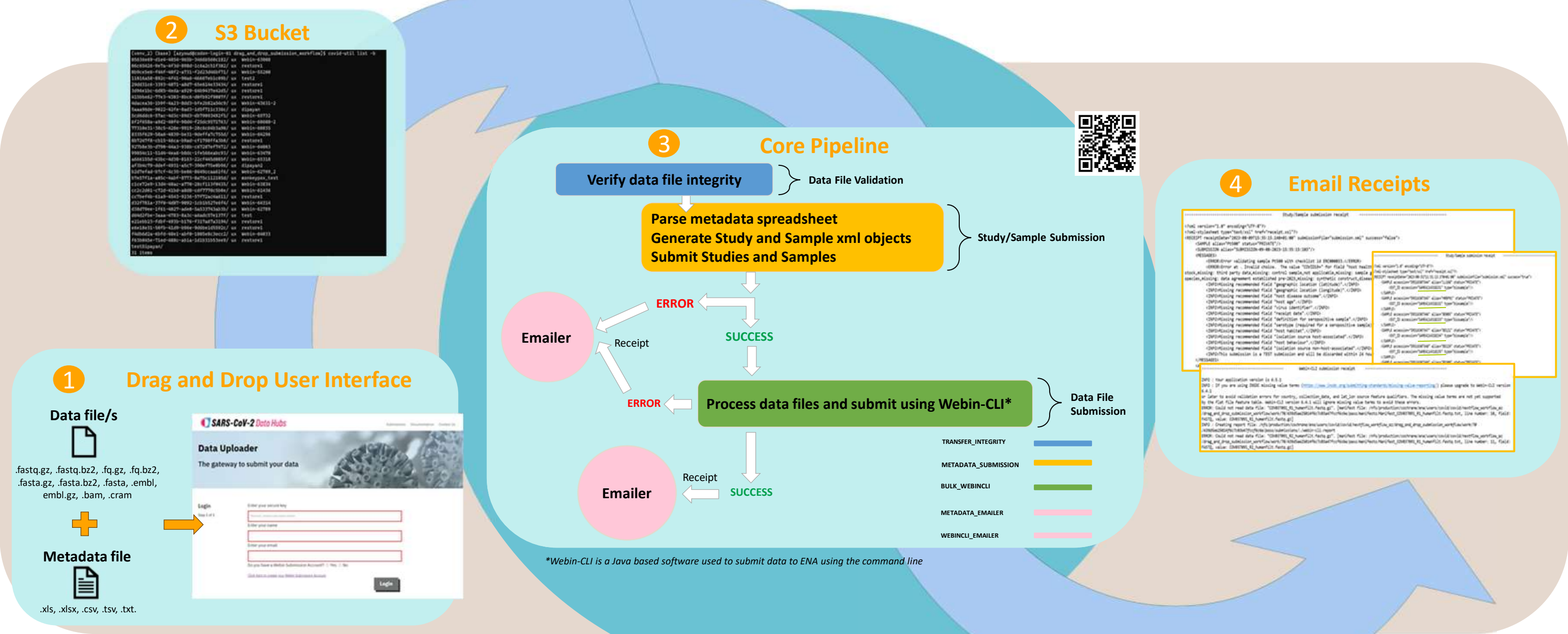
All submissions to the COVID-19 Data Portal must follow the ENA's metadata structure below:



with multiple submission routes possible for each object.

To help maximise the rate and volume of SARS-CoV-2 data shared to the Portal, the Drag and Drop Uploader Tool was developed by EBI (ENA and Archive Infrastructure and Technology (AIT)) teams as a simpler alternative to existing ENA submission services.

It requires no technical skills from users and no prior knowledge of the repository's metadata structure nor submission process.



```
[illegible]
```

```
Study/Sample submission receipt
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="receipt.xsl"?>
<RECEIPT receiptDate="[illegible]" submissionFile="submission.xml" success="false">
     <SAMPLE alias="[illegible]" status="PRIVATE"/>
     <SUBMISSION alias="[illegible]"/>
     <MESSAGES>
          <ERROR>Error validating sample [illegible] with checklist id [illegible].</ERROR>
          <INFO>Missing recommended field "geographic location (latitude)".</INFO>
          <INFO>Missing recommended field "geographic location (longitude)".</INFO>
          <INFO>Missing recommended field "host disease outcome".</INFO>
          <INFO>Missing recommended field "host age".</INFO>
          <INFO>Missing recommended field "virus identifier".</INFO>
          <INFO>Missing recommended field "receipt date".</INFO>
          <INFO>Missing recommended field "definition for seropositive sample".</INFO>
          <INFO>Missing recommended field "host habitat".</INFO>
          <INFO>Missing recommended field "isolation source host-associated".</INFO>
          <INFO>Missing recommended field "host behaviour".</INFO>
          <INFO>Missing recommended field "isolation source non-host-associated".</INFO>
          <INFO>This submission is a TEST submission and will be discarded within 24 hours</INFO>
     </MESSAGES>
```

```
Webin-CLI submission receipt
INFO : Your application version is 6.5.1
[illegible]
```

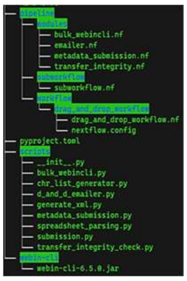
*Webin-CLI is a Java based software used to submit data to ENA using the command line

The SARS-CoV-2 Drag and Drop Nextflow Automation. (1) A user logs in to the interface using secure credentials provided by the ENA, and uploads their data files and a metadata spreadsheet (with the file extensions specified) to a specific Amazon S3 Bucket (2). (3) All data is then transferred to an EBI High Performance Compute environment and submitted through the core pipeline. (4) Notifications of a successful or unsuccessful ENA submission will be received by email.

Pipeline Directory Structure

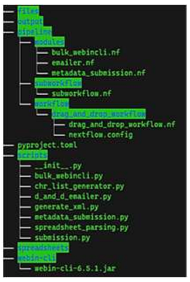
Backend Pipeline

- S3 Bucket transfer
- File integrity validation
- Metadata parsing
- Study and sample metadata submission
- Data file submission
- Email notification



```
├── pipeline
│   ├── modules
│   │   ├── bulk_webincli.nf
│   │   ├── emailer.nf
│   │   ├── metadata_submission.nf
│   │   └── transfer_integrity.nf
│   ├── subworkflow
│   │   └── subworkflow.nf
│   └── workflow
│       └── drag_and_drop_workflow
│           ├── drag_and_drop_workflow.nf
│           └── nextflow.config
├── pyproject.toml
├── scripts
│   ├── __init__.py
│   ├── bulk_webincli.py
│   ├── chr_list_generator.py
│   ├── d_and_d_emailer.py
│   ├── generate_xml.py
│   ├── metadata_submission.py
│   ├── spreadsheet_parsing.py
│   ├── submission.py
│   └── transfer_integrity_check.py
└── webin-cli
    └── webin-cli-6.5.0.jar
```

Standalone Pipeline



```
├── files
├── output
├── pipeline
│   ├── modules
│   │   ├── bulk_webincli.nf
│   │   ├── emailer.nf
│   │   └── metadata_submission.nf
│   ├── subworkflow
│   │   └── subworkflow.nf
│   └── workflow
│       └── drag_and_drop_workflow
│           ├── drag_and_drop_workflow.nf
│           └── nextflow.config
├── pyproject.toml
├── scripts
│   ├── __init__.py
│   ├── bulk_webincli.py
│   ├── chr_list_generator.py
│   ├── d_and_d_emailer.py
│   ├── generate_xml.py
│   ├── metadata_submission.py
│   ├── spreadsheet_parsing.py
│   └── submission.py
├── spreadsheets
└── webin-cli
    └── webin-cli-6.5.1.jar
```

- Metadata parsing
- Study and sample metadata submission
- Data file submission
- Email notification

Limitations

- Workflow does not support 're-runs' - cannot be run from a specific stage in case of errors / partial submissions
- The tool submit only a single data type (Raw reads or genome assemblies) per run
- Currently specific to SARS-CoV-2 data

Feature improvements

- Support re-run of pipeline from specific step, e.g.:
  - Transfer > submit metadata > submit data > emailer
- Submission of both raw reads and genome assemblies in one Nextflow run
- Containerisation with Docker / Singularity
- Refactor hard-coded script variables to support other small pathogen submissions to ENA