



➤ EBAii Assemblage & Annotation 2024

Soumission de données génomiques

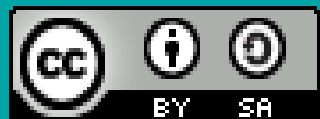
V. Loux



Valentin.loux@inrae.fr

<https://orcid.org/0000-0002-8268-915X>

With contribution from H. Chiapello a,d T. Denecker



•Open Science & FAIR Data



F
indable

Facile à trouver : pour les humains et les ordinateurs :

- id unique et pérennes
- métadonnées riches



A
ccessible

Accessibles à long terme

- entrepôt "pérenne »
- licence d'utilisation explicite (FAIR ≠ ouvert)



I
nteroperable



R
eusable

Interoperables : faciles à combiner avec d'autres jeux de données

- formats ouverts et documentés
- vocabulaire standardisé, ontologies (données et méta-données)

Réutilisables :

- réutilisables par soi, par d'autres
- réutilisables par des machines

Wilkinson 2016

[10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18)

« Digital objects—from data to analytical pipelines—**benefit from application** of these FAIR principles to ensure transparency, reproducibility, and **reusability**. »

Loi pour une République Numérique, 2016 : Obligation de publication des données de la recherche si $\geq 50\%$ financement public

Pour être réutilisable, les données doivent être FAIR

Pour les données de séquence

- « Principes de Bermuda » lors du Human Genome Project (1996) :
 - Automatic release of sequence assemblies larger than 1 kb (preferably within 24 hours).
 - Immediate publication of finished annotated sequences.
 - Aim to make the **entire sequence freely available in the public domain** for both research and development in order to maximise benefits to society.
- Ces principes ne sont **généralement pas suivis sur les aspects** release automatisée et immédiate, **le sont sur la mise à disposition.**
- Progressivement les journaux exigent une mise à disposition des données au moment de la publication
 - Assemblages
 - Données brutes (NGS)

Maxson Jones et al. J Hist Biol. 2018.
INSDC Policy <https://www.insdc.org/policy/>

En pratique ?



INRAE

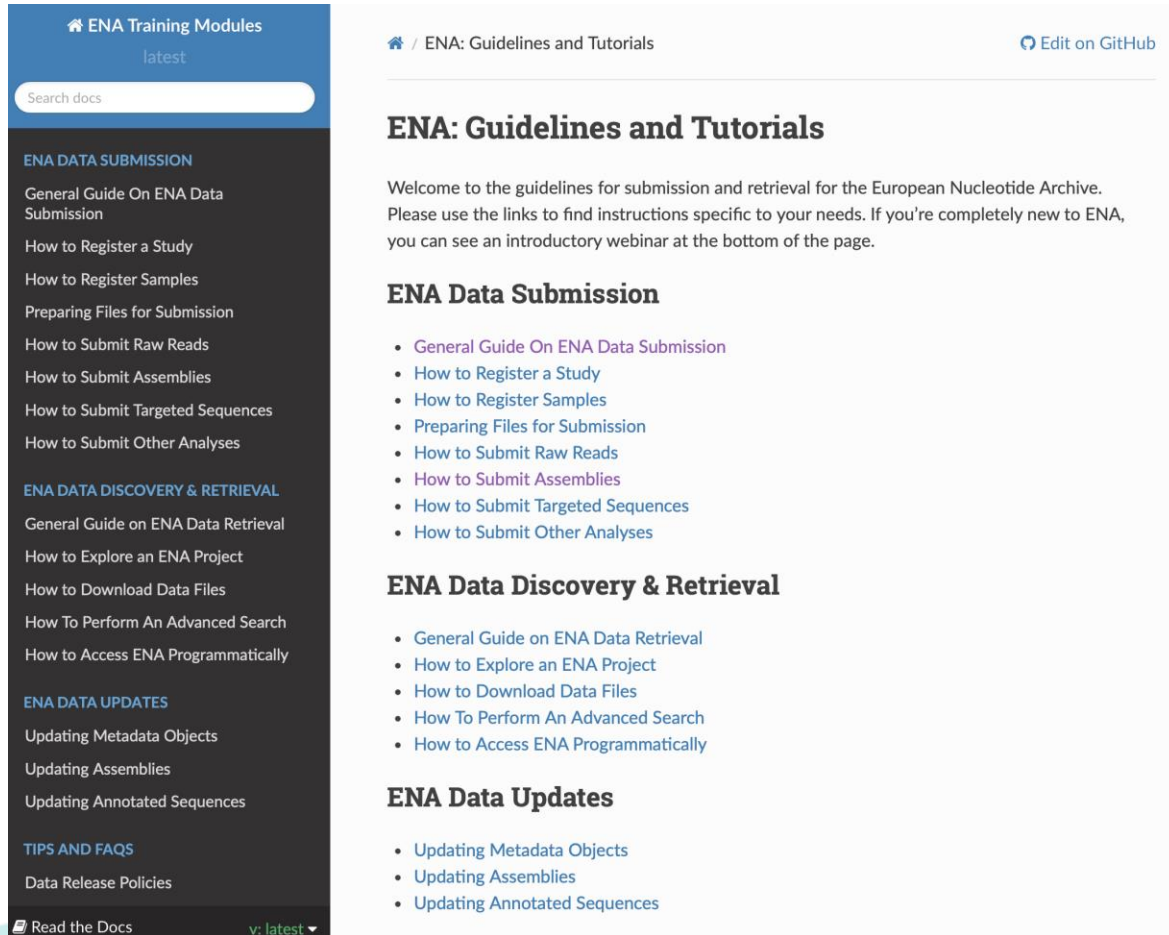
EBAii Assemblage & Annotation

27/09/22/ MaIAGE-Migale/ H. Chiapello, V. Loux

- Entrepôt de données nucléiques
 - Reads
 - Séquences assemblées :
 - Gènes
 - Génomes
 - Métagénomes
- Synchronisé* avec les deux autres banques internationales
 - DDBJ (Japon)
 - NCBI (USA)
- Fournit un ID unique et perenne
- Connecté à UniProt, RNACentral, EBI Metagenomics, Ensembl, Ensembl Genomes, ArrayExpress...

•How – to ?

La bible de la soumission



ENA Training Modules
latest

Search docs

ENA DATA SUBMISSION

- General Guide On ENA Data Submission
- How to Register a Study
- How to Register Samples
- Preparing Files for Submission
- How to Submit Raw Reads
- How to Submit Assemblies
- How to Submit Targeted Sequences
- How to Submit Other Analyses

ENA DATA DISCOVERY & RETRIEVAL

- General Guide on ENA Data Retrieval
- How to Explore an ENA Project
- How to Download Data Files
- How To Perform An Advanced Search
- How to Access ENA Programmatically

ENA DATA UPDATES

- Updating Metadata Objects
- Updating Assemblies
- Updating Annotated Sequences

TIPS AND FAQs

- Data Release Policies

Read the Docs v: latest

ENA: Guidelines and Tutorials Edit on GitHub

ENA: Guidelines and Tutorials

Welcome to the guidelines for submission and retrieval for the European Nucleotide Archive. Please use the links to find instructions specific to your needs. If you're completely new to ENA, you can see an introductory webinar at the bottom of the page.

ENA Data Submission

- General Guide On ENA Data Submission
- How to Register a Study
- How to Register Samples
- Preparing Files for Submission
- How to Submit Raw Reads
- How to Submit Assemblies
- How to Submit Targeted Sequences
- How to Submit Other Analyses

ENA Data Discovery & Retrieval

- General Guide on ENA Data Retrieval
- How to Explore an ENA Project
- How to Download Data Files
- How To Perform An Advanced Search
- How to Access ENA Programmatically

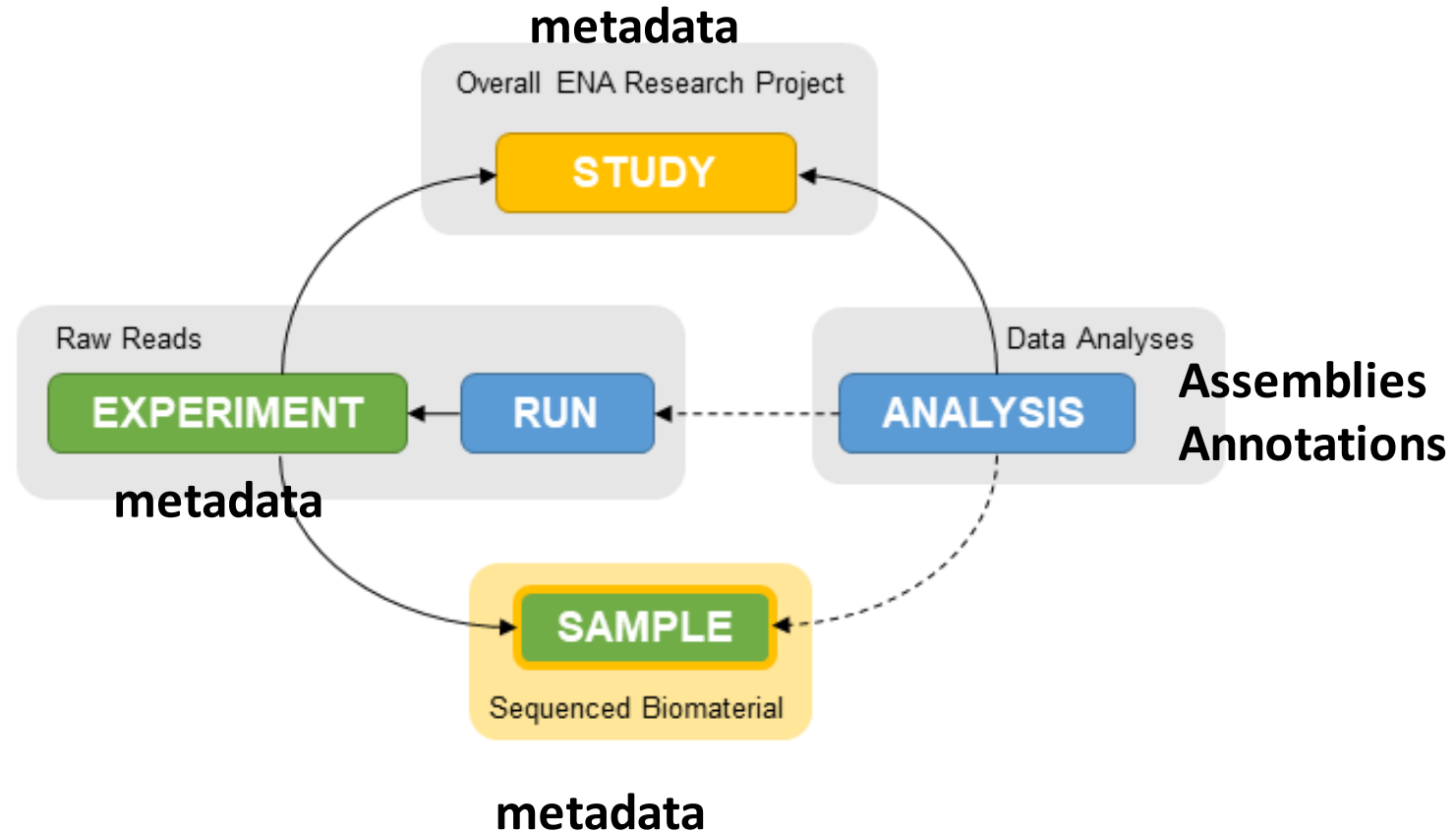
ENA Data Updates

- Updating Metadata Objects
- Updating Assemblies
- Updating Annotated Sequences

<https://ena-docs.readthedocs.io/en/latest/>

Modèle de métadonnées

Tous les
« **samples** »
soumis à ENA
doivent suivre
des
« **checklists** »



•Exemple metadata

Description des experiments

A sequence read submission consists of:

- General experiment information
 - Study accession or unique name (alias)
 - Sample accession or unique name (alias)
 - Experiment name
 - Sequencing platform
 - Sequencing instrument
 - Library name (optional)
 - Library source
 - Library selection
 - Library strategy
 - Free text library description (optional)
 - Insert size for paired reads (optional)
- Read data file(s)
 - BAM file
 - CRAM file
 - Single Fastq file
 - Paired Fastq files
 - Multi-fastq files

Permitted values for platform 🔗

- BGISEQ: Sequencers based on DNBSEQ by MGI Tech.
- CAPILLARY: Sequencers based on capillary electrophoresis technology manufactured by LifeTech (formerly Applied BioSciences).
- DNBSEQ: Uses DNA nanoballs(DNB) and regular array chips.
- ELEMENT: Element Biosciences uses avidity sequencing: a polymerase inserts nucleotides, detected using fluorescence.
- GENAPSYS: Chip based electronic sensing of polymerase extension reaction.
- GENEMIND: Genemind Platform, a single-molecule gene sequencer.
- HELICOS: Helicos is similar to 454 technology - uses 1-color sequential flows.
- ILLUMINA: 54 technology use 1-color sequential flows: 454 technology use 1-color sequential flows.
- ION_TORRENT: Ion Torrent Personal Genome Machine (PGM) from Life Technologies. Directly translates chemically encoded information (A, C, G, T) into digital on semiconductor chip.
- LS454: 454 technology use 1-color sequential flows.
- OXFORD_NANOPORE: Oxford Nanopore platform type. nanopore-based electronic single molecule analysis.
- PACBIO_SMRT: PacificBiosciences platform type for the single molecule real time (SMRT) technology.
- TAPESTRI: Tapestri Platform single-cell DNA genotyping using pre-configured or custom gene panels.
- VELA_DIAGNOSTICS: VELA NGS platform measures the hydrogen ions that are generated during the incorporation of nucleotides in the DNA sequencing reaction.
- ULTIMA: Ultima Genomics platform type. Flowing one nucleotide at a time in order, iteratively.



Meta-data pour « samples »

A checklist defines the minimum and optional metadata expected to describe biological samples

- ENA are based on the Genomic Standards Consortium (GSC) recommendations
- The most suitable checklist depends on the type of the sample:
<https://www.ebi.ac.uk/ena/browser/checklists>
- All ENA checklist are defined by an access number like ERCxxx (Ena R Checklist xxx)
- Example: **ENA prokaryotic pathogen minimal sample checklist**
<https://www.ebi.ac.uk/ena/browser/view/ERC000028>

Sample checklist

Sample Checklists

There is a minimum amount of information required during ENA sample registration and all samples must conform to a defined checklist of expected metadata values. The most suitable checklist for sample registration depends on the type of the sample.

These sample checklists have been developed to meet the needs of different research communities. Different communities have different requirements on the minimum metadata expected to describe biological samples. For any questions regarding checklists, contact us [here](#).

Filter by accession/name/description/field name

Accession	Name	Description
ERC000012	GSC MixS air	Genomic Standards Consortium package extension for reporting of measurements and observations obtained from the environment where the sample was obtained. By choosing the environmental package, a selection of fields can be made from a relevant subsets of the GSC terms.
ERC000013	GSC MixS host associated	Genomic Standards Consortium package extension for reporting of measurements and observations obtained from the environment where the sample was obtained. By choosing the environmental package, a selection of fields can be made from a relevant subsets of the GSC terms.
ERC000014	GSC MixS human associated	Genomic Standards Consortium package extension for reporting of measurements and observations obtained from the environment where the sample was obtained. By choosing the environmental package, a selection of fields can be made from a relevant subsets of the GSC terms.
ERC000015	GSC MixS human gut	Genomic Standards Consortium package extension for reporting of measurements and observations obtained from the environment where the sample was obtained. By choosing the environmental package, a selection of fields can be made from a relevant subsets of the GSC terms.
ERC000016	GSC MixS human oral	Genomic Standards Consortium package extension for reporting of measurements and observations obtained from the environment where the sample was obtained. By choosing the environmental package, a selection of fields can be made from a relevant subsets of the GSC terms.

Submission methods

	Interactive	Webin-CLI	Programmatic
Study	Y	N	Y
Sample	Y	N	Y
Read data	Y	Y	Y
Genome Assembly	N 🤯	Y	N
Transcriptome Assembly	N	Y	N
Template Sequence	N	Y	N
Other Analyses	N	N	Y

Interactive

Dashboard

Welcome to the Webin Submissions Portal

You can use this service for a range of submission activities as well as reports on your submissions. For help with submitting your data, including the use of this interface, please refer to our [Help Guides](#). Please familiarise yourself with the different submission interfaces and what can be submitted through each by reading our [General Guide on ENA Data Submission](#). All users are advised to take a moment to understand the [ENA Metadata Model](#). You may also like to review how the release of data is managed in our [Data Release FAQ](#).

A dedicated submission API for COVID-19 genomes is available [here](#).

Studies (Projects)

- Register Study
- Submit XMLs (advanced)
- Studies Report

Samples

- Register Samples
- Register Novel Taxonomy
- Submit XMLs (advanced)
- Samples Report

Raw Reads (Experiments and Runs)

Raw reads can also be submitted using [Webin-CLI](#)

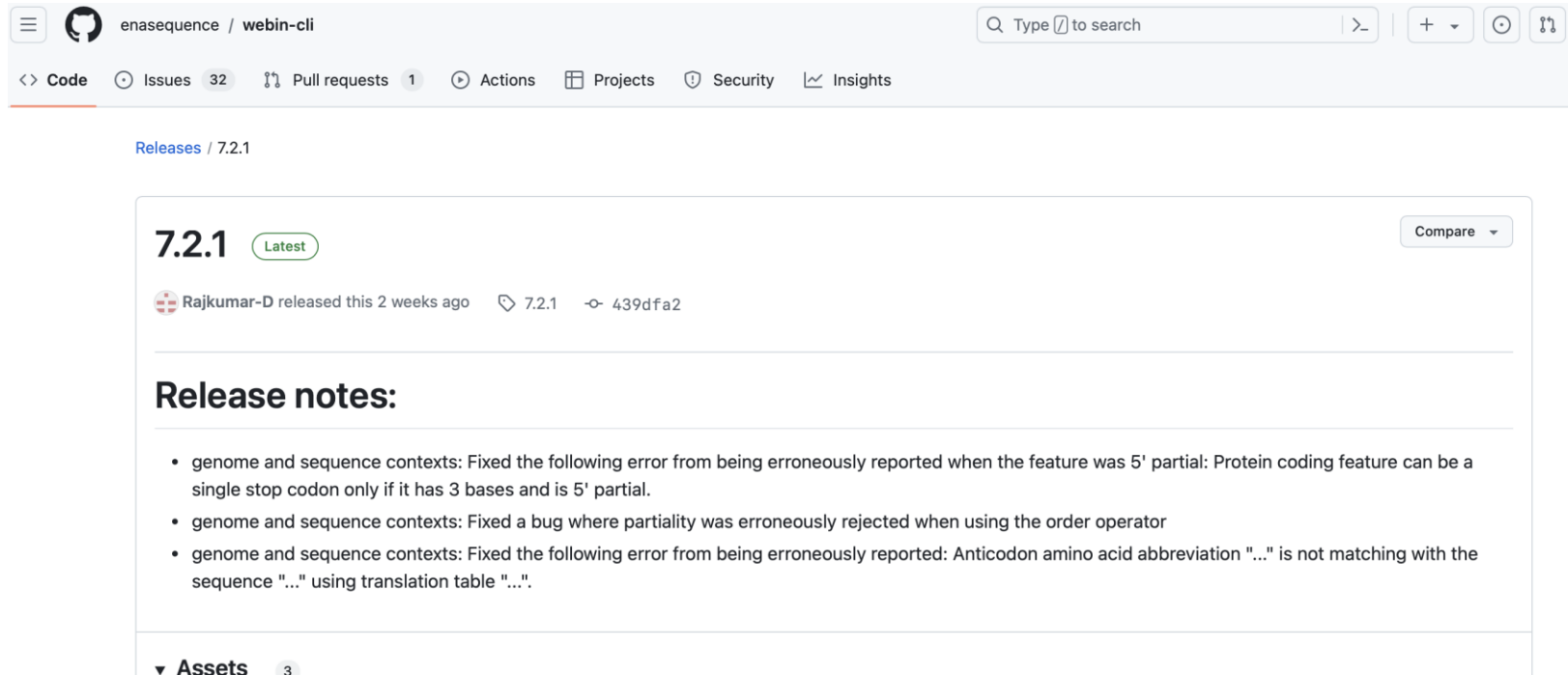
- Submit Reads
- Submit XMLs (advanced)
- Runs Report
- Run Files Report
- Run Processing Report
- Unsubmitted Files Report

Data Analyses

Assemblies and associated sequences must be submitted with [Webin-CLI](#). Other analyses can be submitted as XMLs.

- Generate Annotated Sequence Spreadsheet
- Submit XMLs (advanced)
- Analyses Report
- Analysis File Report
- Analysis Processing Report

Webin-CLI



The screenshot shows the GitHub interface for the repository `enasequence/webin-cli`. The top navigation bar includes links for Code, Issues (32), Pull requests (1), Actions, Projects, Security, and Insights. Below the navigation bar, the 'Releases' section is active, displaying the latest release, 7.2.1, which was released 2 weeks ago by Rajkumar-D. The release notes section lists three updates: fixing an error for 5' partial protein coding features, a bug with the order operator, and an error with the anticodon amino acid abbreviation. The Assets section is partially visible at the bottom.

Using project & sample id :

- Prepare files
- **Validate files !!!!**
- **Submit (upload)**



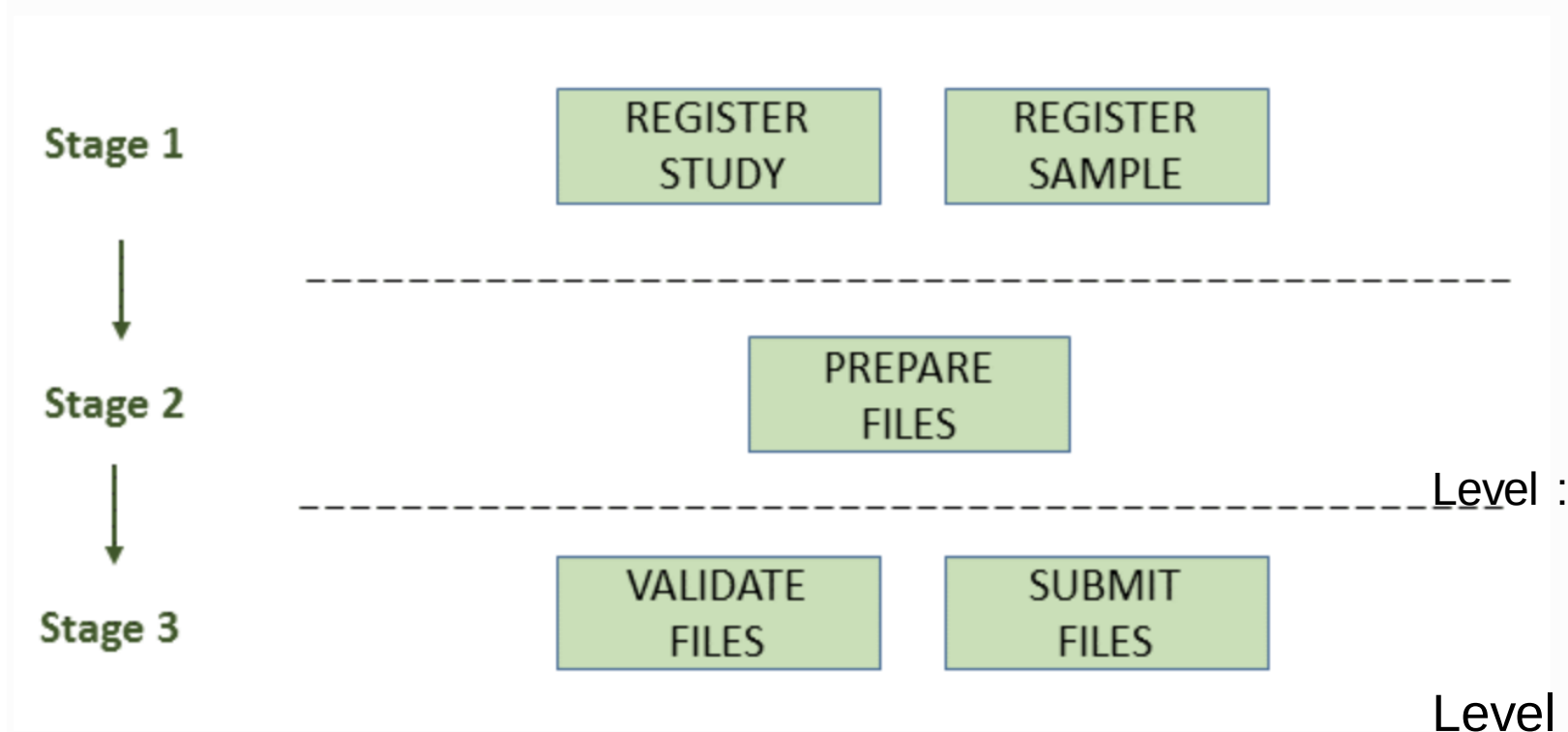
INRAE

EBAii Assemblage & Annotation

27/09/22/ MaIAGE-Migale/ H. Chiapello, V. Loux

<https://github.com/enasequence/webin-cli/releases/tag/7.2.1>

En pratique :



Level : Easy 😎 Validate & submit

Level :
from 😐
to 😱

- Register study & sample
- Don't forget to reserve locus_tag if you want one
- Submit raw reads
 - Prepare files for Webin-CLI
- Submit assembly & annotation
 - Prepare file for Webin-CLI
 - Submit assembly (fasta)
 - Submit annotation (EMBL)

• En pratique

- Ca prend du temps et c'est (malheureusement) complexe
 - A faire au plus tôt (read puis génomes)
 - En gardant les séquences sous embargo (periodes de 2 ans max renouvelables)
- Attention aux critères de filtre de dépôt d'assemblage (ie contigs >200 nct), dès le début de l'analyse.
- Mettre des adresses génériques comme « submitter » (seul autorisé à mettre à jour les données. Au défaut, un permanent.
- Attention aux méta-données, c'est ce qui fera que vos données sont réutilisables

