



DATA
TERRA



SeBiMER
Bioinformatique marine



ODATIS

athENA a tool for sequencing data management

Atelier technique #17

'Données bioinformatiques de diversité'

Pauline Auffret - SeBiMER

Contents

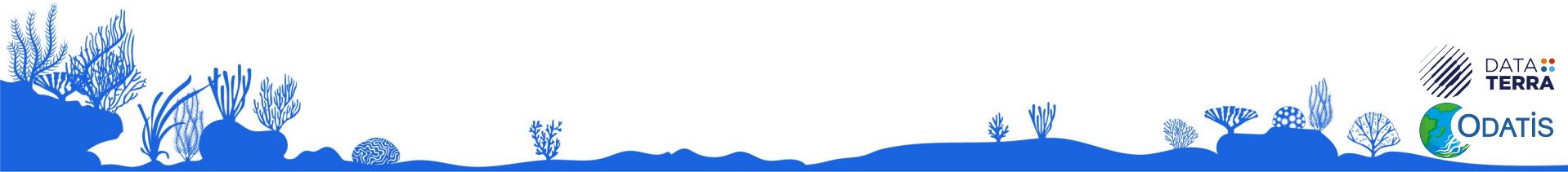
Introduction : be FAIR at Ifremer

Sequencing data metadata

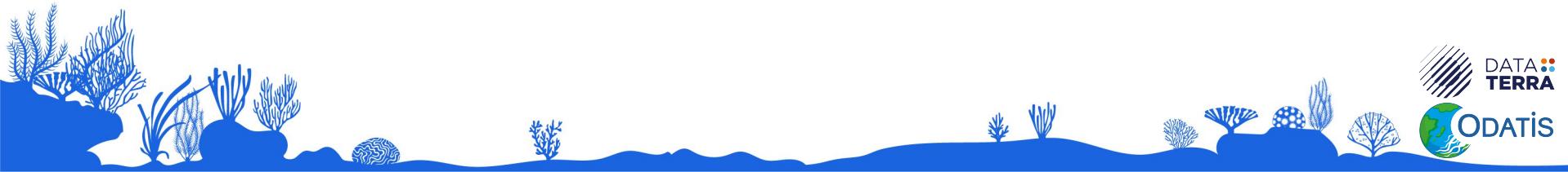
The ENA metadata model

athENA in practice

athENA : what's next ?



Introduction : be F_{indable} A_{ccessible} I_{nteroperable} R_{eusable} at Ifremer



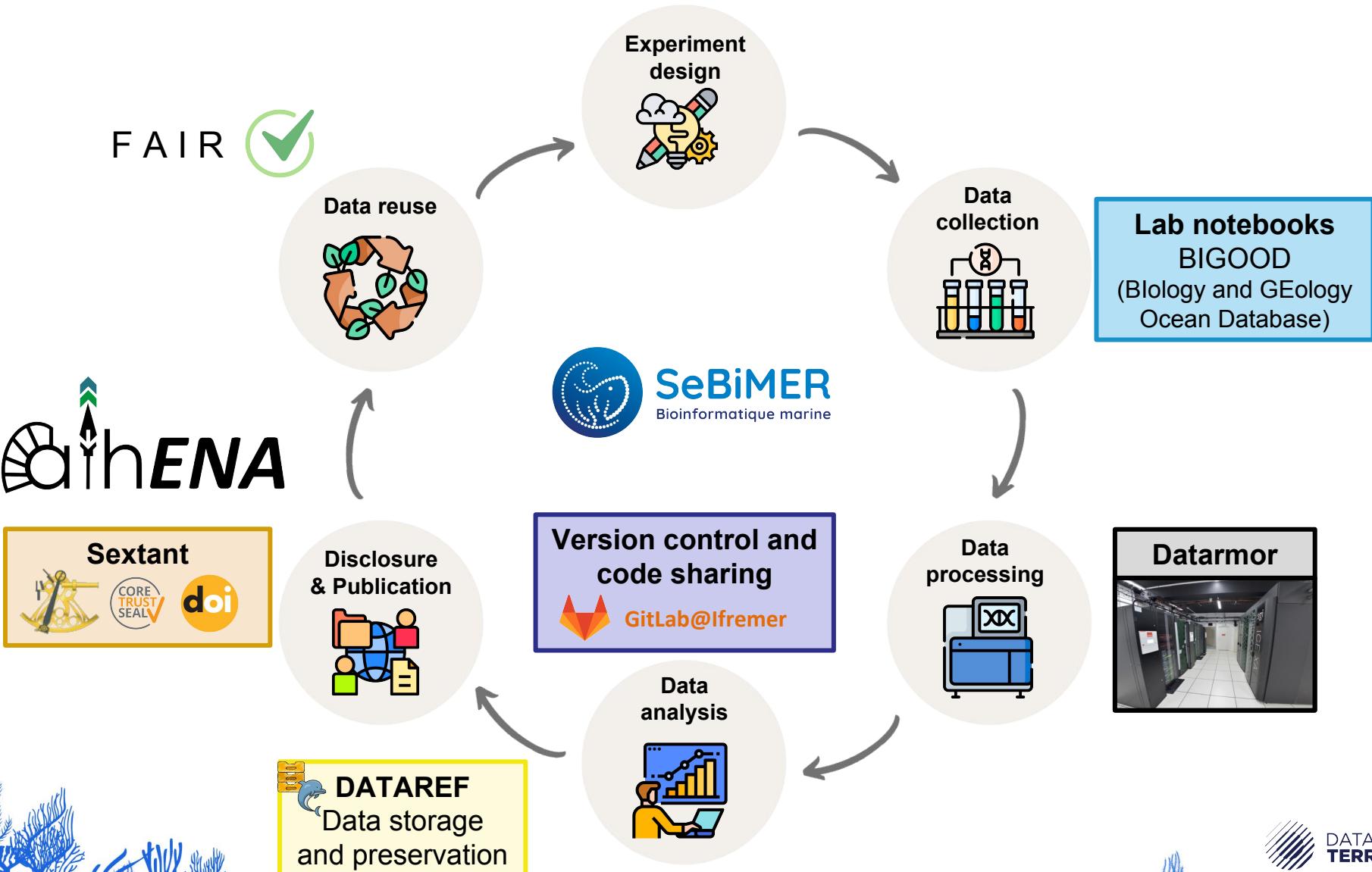
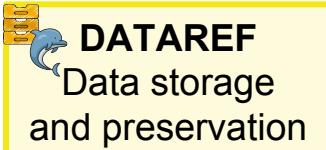
Sequencing data life cycle at Ifremer



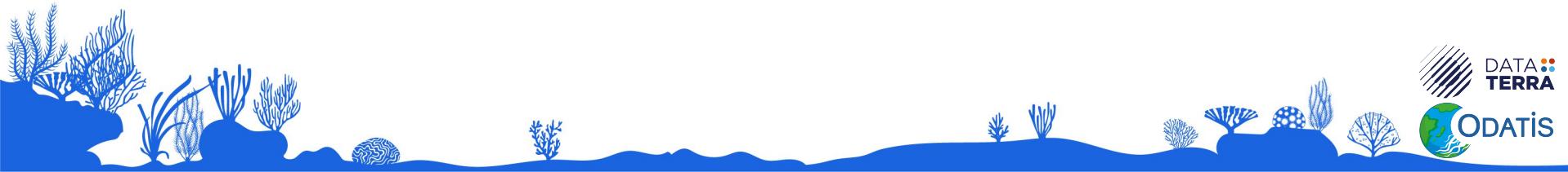
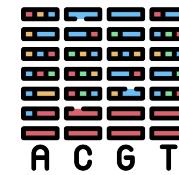
FAIR



bihENA

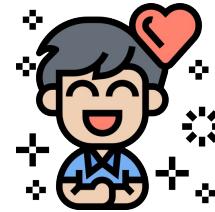


Sequencing data metadata



The importance of metadata

Metadata are **data about the data**



WHAT
WHO
WHEN
WHERE
HOW
WHY

...

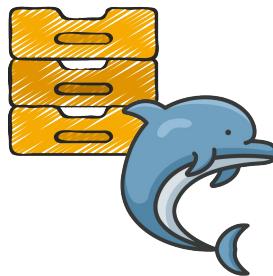
1. Example inspired by F. de Lamotte, IFB, *Introduction aux métadonnées*. https://ifb-elixirfr.github.io/IFB-FAIR-data-training/sequences/module3_sequence1_edition1_cours.html (Accessed on March 11, 2024)

What is **athENA** ?

athENA is a sequencing (meta)data management and brokering tool.

Three main goals :

- metadata collection in compliance with ENA metadata model ;
- sequencing data brokering to ENA database ;
- international indexing for Ifremer' sequencing data.

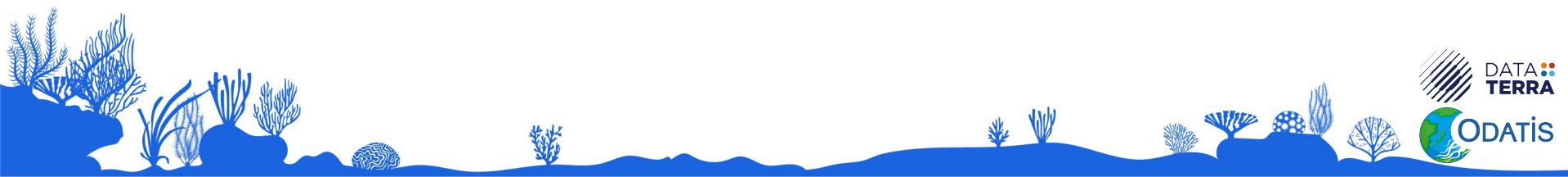


Local storage
(e.g. Dataref at Ifremer)



Metadata collection
Data brokering

Publicly available in
international nucleotide
databases



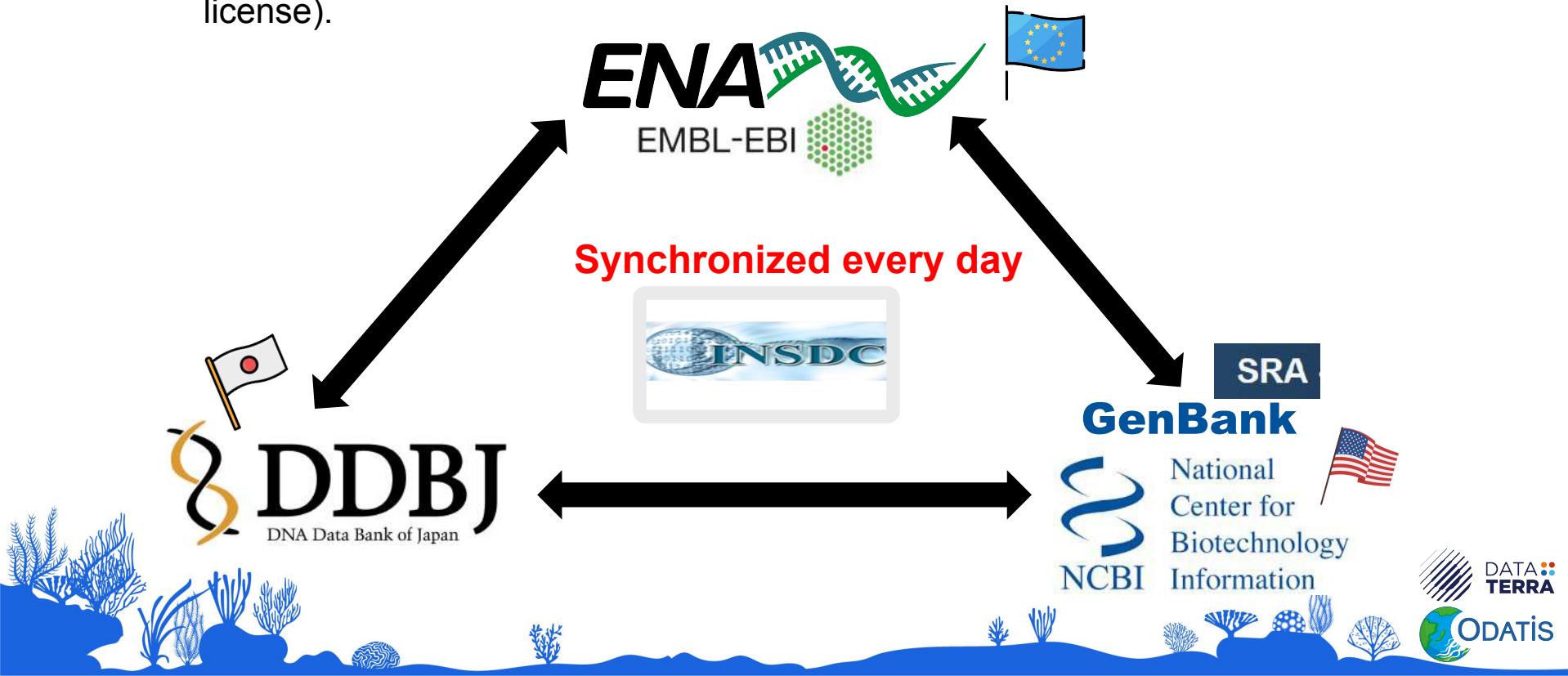
International referencing



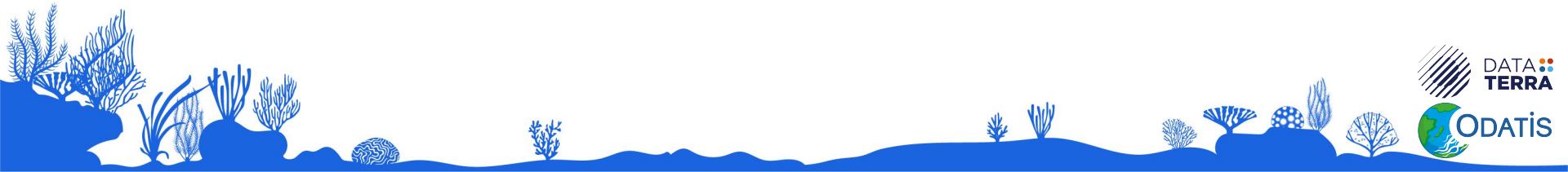
The International Nucleotide Sequence Database Collaboration (INSDC) collects and disseminates international databases containing DNA and RNA sequences.



All data is **available** for free and unrestricted access, for any purpose, with no restrictions on analysis, redistribution, or re-publication of the data (CC-BY license).



The ENA metadata model

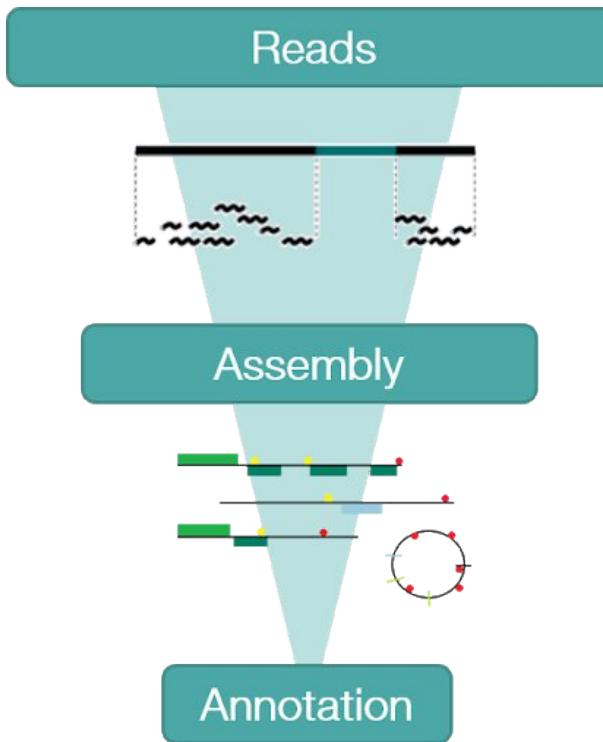


ENA : European Nucleotide Archive

ENA is based on the following principles :

- Three tiers data structure
- 6-object metadata model
- Sample metadata checklists
- Communities ontologies

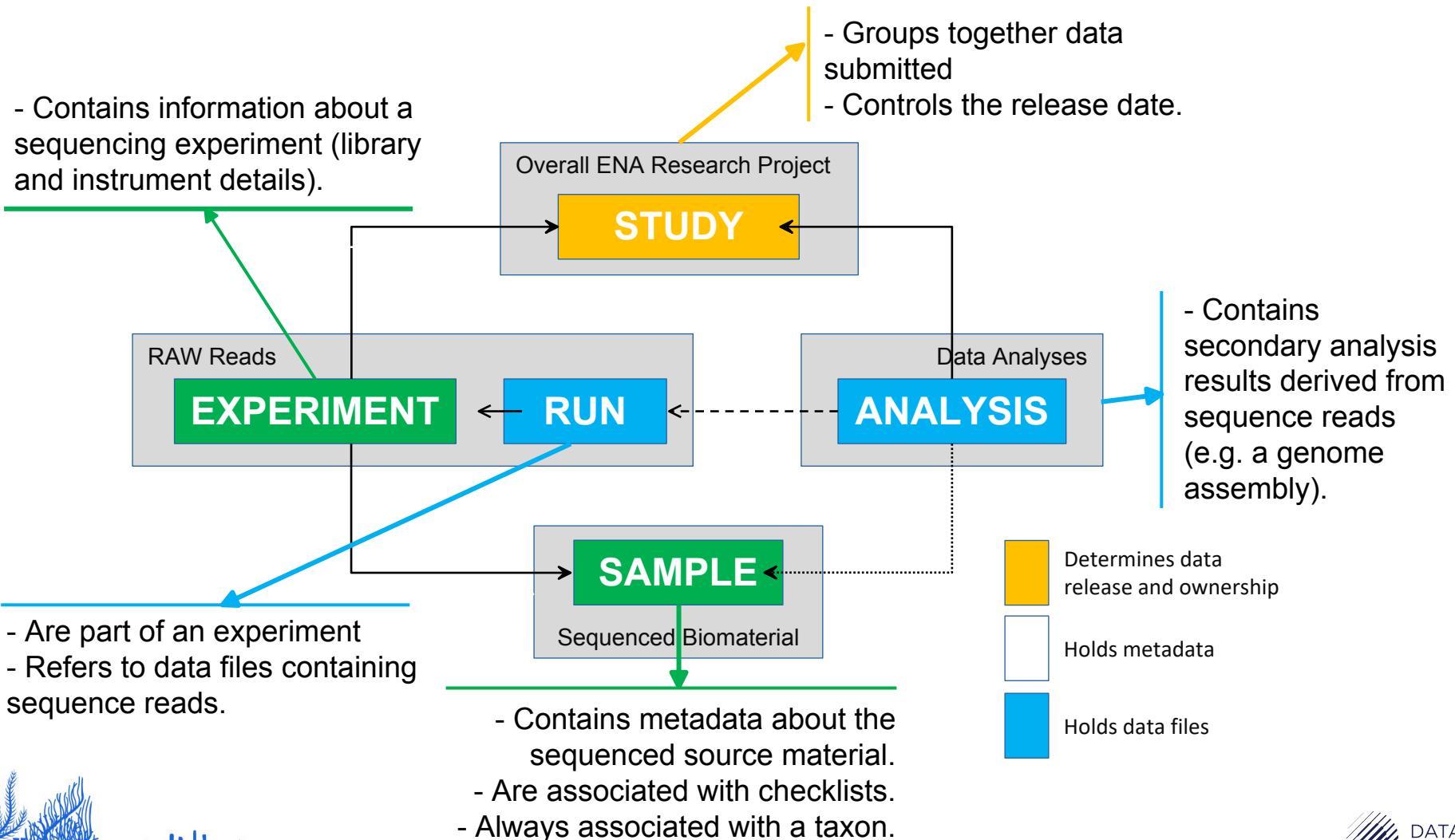
ENA 3 tiers data structure



<https://www.ebi.ac.uk/training/online/courses/ena-quick-tour/what-is-ena/>



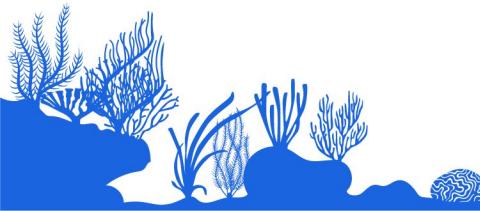
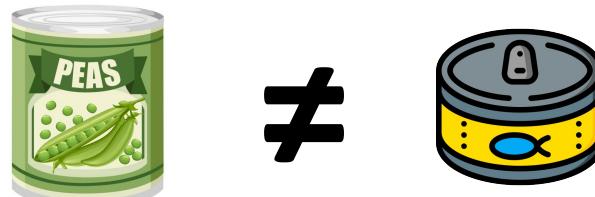
ENA 6-object data structure



ENA sample checklists

ENA sample checklists ensure that a minimum amount of information is provided.

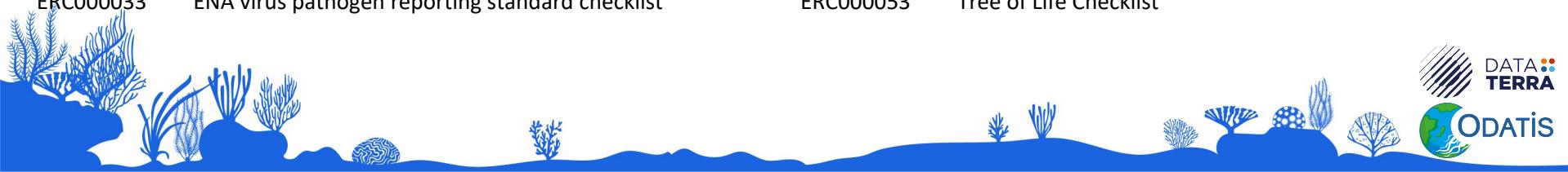
- There is a **minimum amount of information** required during ENA sample registration and all samples must conform to a **defined checklist of expected metadata values**.
- These sample checklists have been developed to **meet the needs of different research communities**. Different communities have different requirements on the minimum metadata expected to describe biological samples.



ENA sample checklists

14 pre-selected checklists for Ifremer projects

Accession	Name	Accession	Name
ERC000011	ENA default sample checklist	ERC000034	ENA mutagenesis by carcinogen treatment checklist
ERC000012	GSC MLxS air	ERC000035	ENA Crop Plant sample enhanced annotation checklist
ERC000013	GSC MLxS host associated	ERC000036	ENA sewage checklist
ERC000014	GSC MLxS human associated	ERC000037	ENA Plant Sample Checklist
ERC000015	GSC MLxS human gut	ERC000038	ENA Shellfish Checklist
ERC000016	GSC MLxS human oral	ERC000039	ENA parasite sample checklist
ERC000017	GSC MLxS human skin	ERC000040	ENA UniEuk_EukBank Checklist
ERC000018	GSC MLxS human vaginal	ERC000041	ENA Global Microbial Identifier Proficiency Test (GMI PT) checklist
ERC000019	GSC MLxS microbial mat biofilm	ERC000043	ENA Marine Microalgae Checklist
ERC000020	GSC MLxS plant associated	ERC000044	COMPARE-ECDC-EFSA pilot human-associated reporting standard
ERC000021	GSC MLxS sediment	ERC000045	COMPARE-ECDC-EFSA pilot food-associated reporting standard
ERC000022	GSC MLxS soil	ERC000047	GSC MIMAGS (Minimum Information about a Metagenome-Assembled Genome)
ERC000023	GSC MLxS wastewater sludge	ERC000048	GSC MISAGS (Minimum Information about a Single Amplified Genome)
ERC000024	GSC MLxS water	ERC000049	GSC MIUVIGS (Minimum Information about an Uncultivated Virus Genome)
ERC000025	GSC MLxS miscellaneous natural or artificial environment	ERC000050	ENA binned metagenome
ERC000027	ENA Micro B3	ERC000051	PDX Checklist
ERC000028	ENA prokaryotic pathogen minimal sample checklist	ERC000052	HoloFood Checklist
ERC000029	ENA Global Microbial Identifier reporting standard checklist GMI_MDM:1.1	ERC000053	Tree of Life Checklist
ERC000030	ENA Tara Oceans		
ERC000031	GSC MLxS built environment		
ERC000032	ENA Influenza virus reporting standard checklist		
ERC000033	ENA virus pathogen reporting standard checklist		



ENA sample checklists

14 pre-selected checklists for Ifremer projects

Accession	Name	Accession	Name
ERC000011	ENA default sample checklist	ERC000034	ENA mutagenesis by carcinogen treatment checklist
ERC000012	GSC MIxS air	ERC000035	ENA Crop Plant sample enhanced annotation checklist
ERC000013	GSC MIxS host associated	ERC000036	ENA sewage checklist
ERC000014	GSC MIxS human associated	ERC000037	ENA Plant Sample Checklist
ERC000015	GSC MIxS human gut	ERC000038	ENA Shellfish Checklist
ERC000016	GSC MIxS human oral	ERC000039	ENA parasite sample checklist
ERC000017	GSC MIxS human skin	ERC000040	ENA UniEuk_EukBank Checklist
ERC000018	GSC MIxS human vaginal	ERC000041	ENA Global Microbial Identifier Proficiency Test (GMI PT) checklist
ERC000019	GSC MIxS microbial mat biofilm	ERC000043	ENA Marine Microalgae Checklist
ERC000020	GSC MIxS plant associated	ERC000044	COMPARE-ECDC-EFSA pilot human-associated reporting standard
ERC000021	GSC MIxS sediment	ERC000045	COMPARE-ECDC-EFSA pilot food-associated reporting standard
ERC000022	GSC MIxS soil	ERC000047	GSC MIMAGS (Minimum Information about a Metagenome-Assembled Genome)
ERC000023	GSC MIxS wastewater sludge	ERC000048	GSC MISAGS (Minimum Information about a Single Amplified Genome)
ERC000024	GSC MIxS water	ERC000049	GSC MIUVIGS (Minimum Information about an Uncultivated Virus Genome)
ERC000025	GSC MIxS miscellaneous natural or artificial environment	ERC000050	ENA binned metagenome
ERC000027	ENA Micro B3	ERC000051	PDX Checklist
ERC000028	ENA prokaryotic pathogen minimal sample checklist	ERC000052	HoloFood Checklist
ERC000029	ENA Global Microbial Identifier reporting standard checklist GMI_MDM:1.1	ERC000053	Tree of Life Checklist
ERC000030	ENA Tara Oceans		
ERC000031	GSC MIxS built environment		
ERC000032	ENA Influenza virus reporting standard checklist		
ERC000033	ENA virus pathogen reporting standard checklist		



ENA sample checklists

Accession Name

ERC000011 ENA default sample checklist



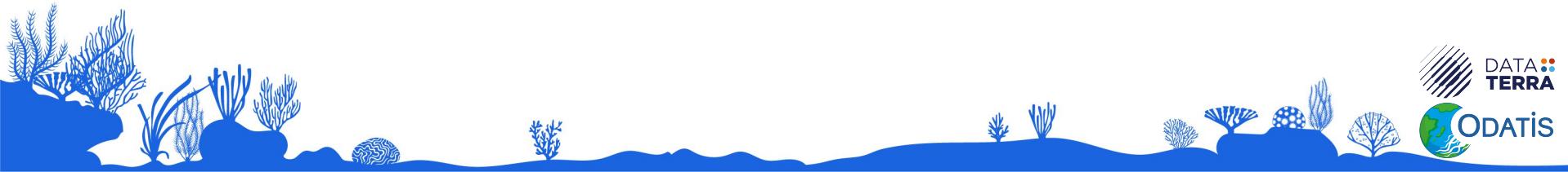
In practice, we often use the default sample checklist because of lack of existing checklists adapted specifically to ifremer projects.



An update is scheduled for March 15, 2024 :

4 new MIxS checklists have been added to ENA:

- GSC MIxS Agriculture
- GSC MIxS Food and Production
- GSC MIxS Symbiont
- GSC MIxS Hydrocarbon



Ontologies

Ontologies are interoperable, logically well-defined, machine readable controlled vocabularies

Ontology is a **structured set of terms and concepts** of a particular domain specifying the **relationships** between these terms and their properties.

Each term of an ontology must have a definition to be sure of the associated meaning.

An ontology presents a **hierarchical structure** and the set of terms is anchored by a high-level term, the root.



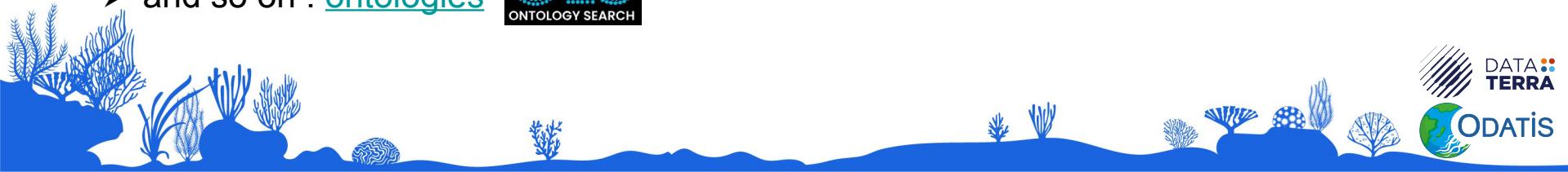
Ontologies

Examples :

- The **GENEONTOLOGY (GO)** aims to maintain and develop controlled vocabulary of **gene and gene product attributes**

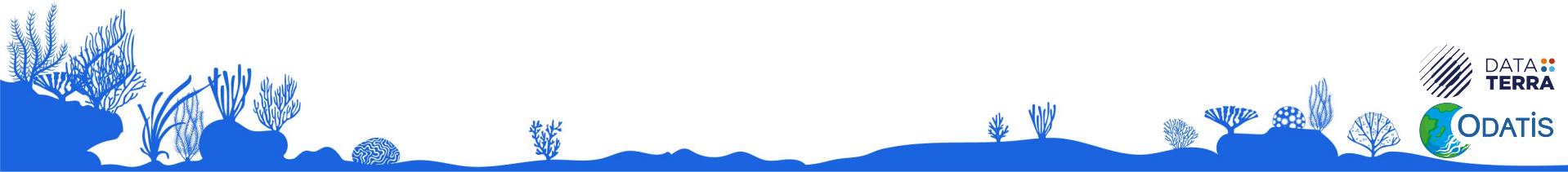


- **EDAM** is an ontology of well established, familiar concepts that are prevalent within bioinformatics, including **types of data and data identifiers, data formats, operations and topics**.
- **ENVO** is an ontology which represents knowledge about **environments, environmental processes, ecosystems, habitats, and related entities**
- and so on : [ontologies](#)



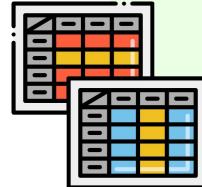


aihENA in practice



aihENA in a nutshell

nextflow



Metadata collection

- User friendly **Excel** sheets
- Drop-down lists with controlled vocabulary
- ENA standards



Metadata validation

- Python pipeline to convert Excel sheets to XML files
- Metadata validation using controlled vocabulary, checklists requirements and format

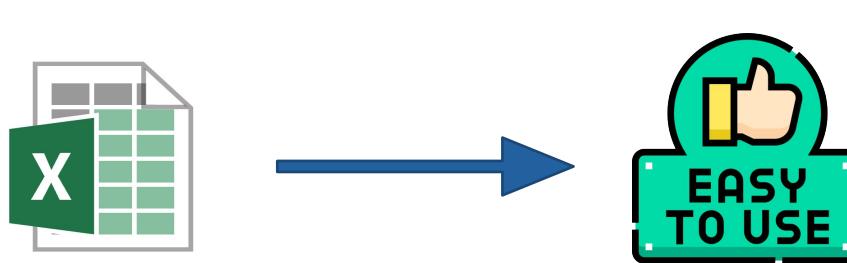


ENA submission

- Data upload to ENA server using cURL
- Embargo max 2 years
- Accession IDs delivered as soon as data are uploaded

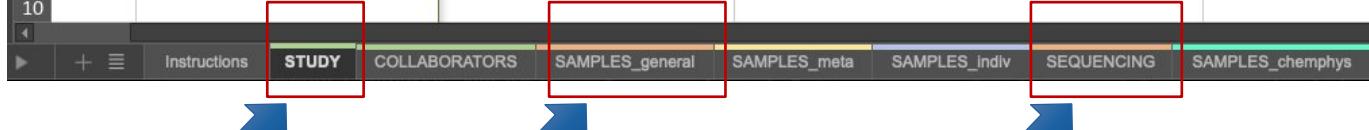
Metadata collection via an Excel file

- **content** : 7 editable sheets corresponding to ENA metadata model
- **rules** : mandatory / optional fields according to selected checklists
- **format** : free text / controlled vocabulary / drop-down lists
- **joker values for mandatory fields** : “not collected” ; “not provided”, “not applicable” ; “restricted access”



Metadata collection via an Excel file

A	B	C	D	E
1 rule	mandatory	mandatory	mandatory	mandatory
2 format	free text	free text	free text	free text
3 help	Temporary accession number, format : YYYYMMDD_LABNAME_SUBMIT TER-INITIALS	Short name for the study (e.g. DATAREF project name)	Title of the study as would be used in a publication. Must contain the following elements : study type, genus, species, project name, year. Example : Whole genome sequencing of Atlantic bluefin tuna for THON project, 2022.	More extensive free description of the st
4 tag	alias	name	title	study_description
5 value				
6				
7				
8				
9				
10				



Overall ENA Research Project

STUDY

SAMPLE

Sequenced Biomaterial

RAW Reads

EXPERIMENT

RUN

Data Analyses

ANALYSIS

mandatory

mandatory

mandatory

optional

Metadata collection via an Excel file

A	B	C	D	E
1 rule	mandatory	mandatory	mandatory	mandatory
2 format	free text	free text	free text	free text
3 help	Temporary accession number, format : YYYYMMDD_LABNAME_SUBMIT TER-INITIALS	Short name for the study (e.g. DATAREF project name)	Title of the study as would be used in a publication. Must contain the following elements : study type, genus, species, project name, year. Example : Whole genome sequencing of Atlantic bluefin tuna for THON project, 2022.	More extensive free description of the st
4 tag	alias	name	title	study_description
5 value				
6				
7				
8				
9				
10				



List here owner
of data

STUDY

mandatory



ENA mandatory
fields only for
metaB projects

SAMPLE



Optional metadata when applied:
- description of individuals
- physico-chemical information

SAMPLE

optional

Metadata collection via an Excel file

scientific_name	common_name	taxon_id	collection_date	isolation_source
seawater metagenome	seawater metagenome	1561972	2020-11-23	control sample
seawater metagenome	seawater metagenome	1561972	2020-11-23	control sample
seawater metagenome	seawater metagenome	1561972	2020-09-03	coastal sea water

sample salinity	sample temperature
39.2	24.7
38.8	16.0
39.9	13.4

country	locality	latitude	longitude
France	Baie des Veys	49.3258	-1.1127
France	Etang de Thau	43.4561	3.6723
France	Etang de Thau	43.4561	3.6723

environment (biom)	environment (feature)	environment (material)	target gene	target subfragment
ocean	not applicable	coastal sea water	16S rRNA	V4-V5
ocean	not applicable	coastal sea water	16S rRNA	V4-V5
sea	mediterranean sea	sea water	16S rRNA	V4-V5

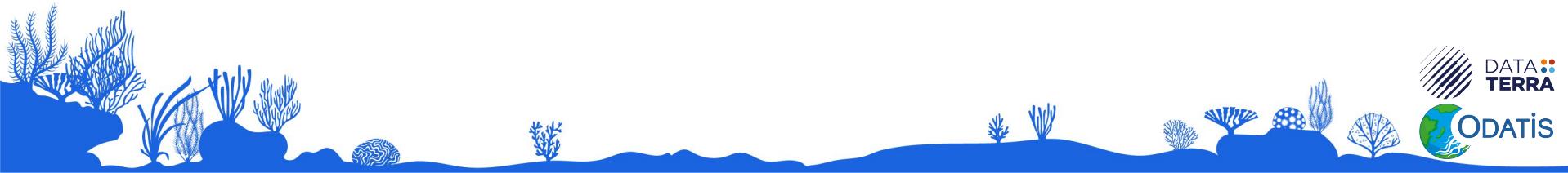
PLATFORM	FILETYPE	pcr primers	multiplex id adapters	
ILLUMINA	fastq	F:GTGYCAGCMGCCGCGTAA-R:CCGYCAATTYMTTTRAGTTT	AGTTTG	F:CTTCCCTACACGACGCTTCCGATCT-R:GGAGTTCA
ILLUMINA	fastq	F:GTGYCAGCMGCCGCGTAA-R:CCGYCAATTYMTTTRAGTTT	GGACGG	F:CTTCCCTACACGACGCTTCCGATCT-R:GGAGTTCA
ILLUMINA	fastq	F:GTGYCAGCMGCCGCGTAA-R:CCGYCAATTYMTTTRAGTTT	TCAGCG	F:CTTCCCTACACGACGCTTCCGATCT-R:GGAGTTCA



ENA programmatic submission

3 routes for ENA submissions, all appropriate, maybe complementary :

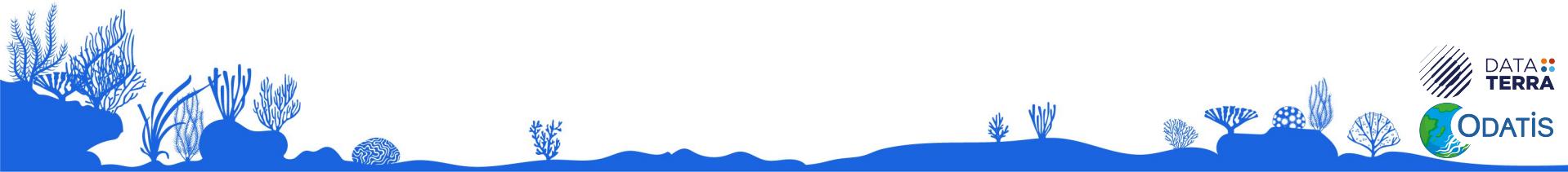
- **Interactive Submissions** : filling out by web forms or off-line downloaded template spreadsheets and uploaded to ENA. Most accessible submission route.
- **Command Line Submissions** : Webin-CLI program. Validates submissions. Allow you maximum control of the process.
- **Programmatic Submissions** : data must be in XML format. Send to ENA by Webin Portal or cURL.



ENA programmatic submission

	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

athENA uses the programmatic route to submit raw reads, and Webin-CLI for genomes.



ENA published data with athENA

The screenshot shows the ENA homepage with a teal header. The header includes the ENA logo, a search bar with placeholder "Enter text search terms" and a "Search" button, and a "View" button next to a search result "PRJEB59362". Below the header are navigation links: Home, Submit, Search, Rulespace, About, and Support.

Project: PRJEB59362

STUDY

Transgenerational exposure to ocean acidification impacts the hepatic transcriptome of European sea bass

Secondary Study

ERP144411

Accession:

Study Title:

RNAseq analysis reveals that transgenerational exposure to ocean acidification impacts the hepatic transcriptome of European sea bass (*Dicentrarchus labrax*) [Show Less](#)

Center Name:

Ifremer

Study Name:

LIVACID

Broker Name:

SeBiMER

COLLABORATORS:

Mazurais David, Servili Arianna, Mouchel Olivier, Zambonino Jose

INSTITUTE_NAME:

IFREMER_RBE_PHYTNESS

CENTER_PROJECT_NAME:

LIVACID

STUDY_TYPE:

RNASeq

STUDY_ABSTRACT:

Physiological effects of ocean acidification associated to elevated CO₂ concentrations in seawater is the subject of numerous studies in teleost fish. While short time impact of ocean acidification on acid-base exchange and energy metabolism are relatively well described, the longer-term effects are much less known. In this study, we investigated the effect of transgenerational exposure to ocean acidification on the hepatic transcriptome of European sea bass (*Dicentrarchus labrax*). [Show Less](#)

ENA-FIRST-PUBLIC:

2023-02-09

ENA-LAST-UPDATE:

2023-02-09

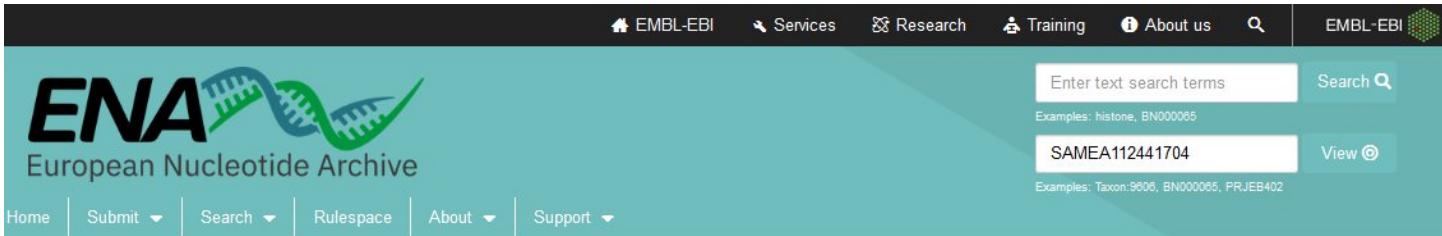


ENA published data with athENA

The screenshot shows the ENA website interface. At the top, there's a navigation bar with links to EMBL-EBI, Services, Research, Training, About us, and a search bar. Below the navigation is the ENA logo and the text "European Nucleotide Archive". A search bar at the top right contains the text "SAMEA112441704" and a "View" button. The main content area displays a biosample record for "SAMEA112441704". The record includes fields like Organism, Sample Accession, Sample Title, Center Name, Sample Alias, Checklist, Broker Name, ENA-CHECKLIST, Collection Date, Sample Description, Environmental Sample, Aquaculture Origin, Geographic Location (Longitude), and Geographic Location (Latitude). The sample title is "97_Liv". A green box on the left labeled "SAMPLE" highlights the record.

Organism:	Dicentrarchus labrax (European seabass);Dicentrarchus labrax
Sample Accession:	SAMEA112441704
Sample Title:	97_Liv
Center Name:	Ifremer
Sample Alias:	sam_97_S5
Checklist:	ERC0000011
Broker Name:	Bioinformatics Core Facility of Ifremer (French Research Institute for Exploitation of the Sea)
ENA-CHECKLIST:	ERC0000011
Collection Date:	2019-10-15
Sample Description:	Acid condition sampling day 1
Environmental Sample:	No
Aquaculture Origin:	AOAR (Aquaculture Origin Aquaculture Raised)
Geographic Location (Longitude):	not provided
Geographic Location (Latitude):	not provided

ENA published data with athENA

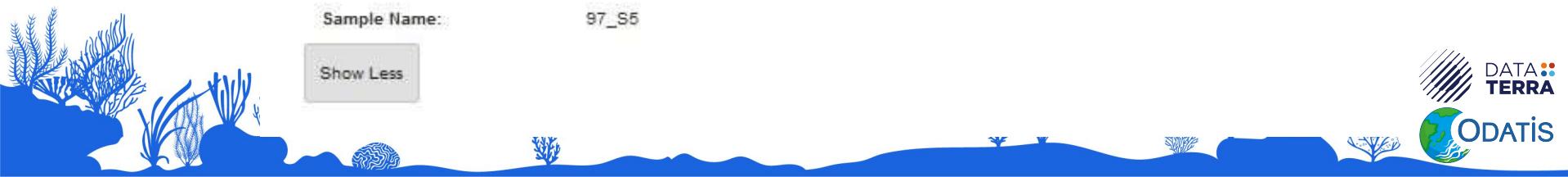


The screenshot shows the ENA homepage. At the top, there's a navigation bar with links to EMBL-EBI, Services, Research, Training, About us, and a search bar. Below the navigation is the ENA logo and the text "European Nucleotide Archive". A search bar contains the text "SAMEA112441704" and a "View" button. Below the search bar, there's a note about examples: "Examples: histone, BN000065" and "Examples: Taxon:9606, BN000065, PRJEB402". At the bottom of the header, there are links for Home, Submit, Search, Rulespace, About, and Support.

SAMPLE

Scientific Name:	Dicentrarchus labrax
Common Name:	European seabass
Dev Stage:	Juvenile
Geographic Location (Region And Locality):	experimental structure - Brest
Investigation Type:	eukaryote
LIBRARY_CONSTRUCTION_PROTOCOL:	Truseq stranded mRNA sample prep kit (Illumina, San Diego, CA, USA)
ENA-FIRST-PUBLIC:	2023-02-09
Project Name:	LIVACID
Sequencing Method:	Sequencing by synthesis (Illumina)
ENA-LAST-UPDATE:	2023-02-09
pH:	7.6
Tissue Type:	Liver
Age:	18 months
Geographic Location (Country And/or Sea):	France
Isolation Source:	liver
Sample Name:	97_S6

Show Less



DATA
TERRA

ENA published data with athENA

ENAL European Nucleotide Archive

Home | Submit | Search | Rulespace | About | Support

Enter text search terms Search

Examples: histone, BN000065

ERX10263838

Examples: Taxon:9606, BN000065, PRJEB402

EXPERIMENT

Experiment: ERX10263838

Illumina NovaSeq 6000 sequencing; 73_S1

Organism:	Dicentrarchus labrax (European seabass)
Experiment Accession:	ERX10263838
Instrument Platform:	ILLUMINA
Instrument Model:	Illumina NovaSeq 6000
Center Name:	Ifremer
Library Layout:	SINGLE
Library Strategy:	RNA-Seq
Library Source:	TRANSCRIPTOMIC
Library Name:	73_S1
Library Selection:	cDNA_oligo_dT
Broker Name:	SeBiMER
LIBRARY_CONSTRUCTION_PROTOCOL:	Truseq stranded mRNA sample prep kit (Illumina, San Diego, CA, USA)

| 50



ENA published data with athENA

RUN

The screenshot shows the ENA homepage with a search bar and navigation links. Below the header, the ENA logo and "European Nucleotide Archive" are displayed. A blue banner at the top says "Run: ERR10818349". The main content area shows the following details for the run:

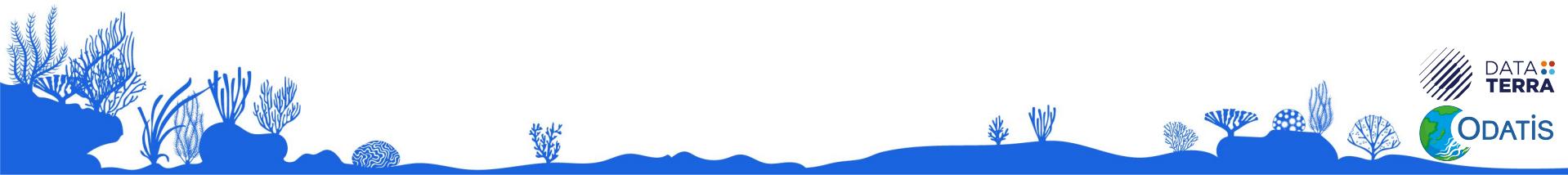
Organism:	Dicentrarchus labrax (European seabass)
Instrument Platform:	ILLUMINA
Instrument Model:	Illumina NovaSeq 6000
Read Count:	75677959
Base Count:	7567795900
Center Name:	Ifremer
Library Layout:	SINGLE
Library Strategy:	RNA-Seq
Library Source:	TRANSCRIPTOMIC
Library Name:	73_S1

Below the table is a "Show More" button. At the bottom, there are download options for "Download report: JSON TSV" and "Get download script Download selected files". A table for "Generated FASTQ files: FTP" lists the study accession, sample accession, experiment accession, run accession, tax id, scientific name, and a checkbox for the fastq.gz file. The checkbox is checked, and the file name is "ERR10818349.fastq.gz".



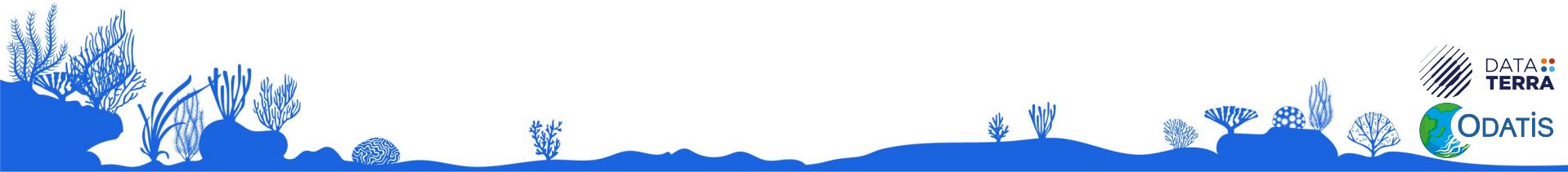


OAIhENA : what's next ?

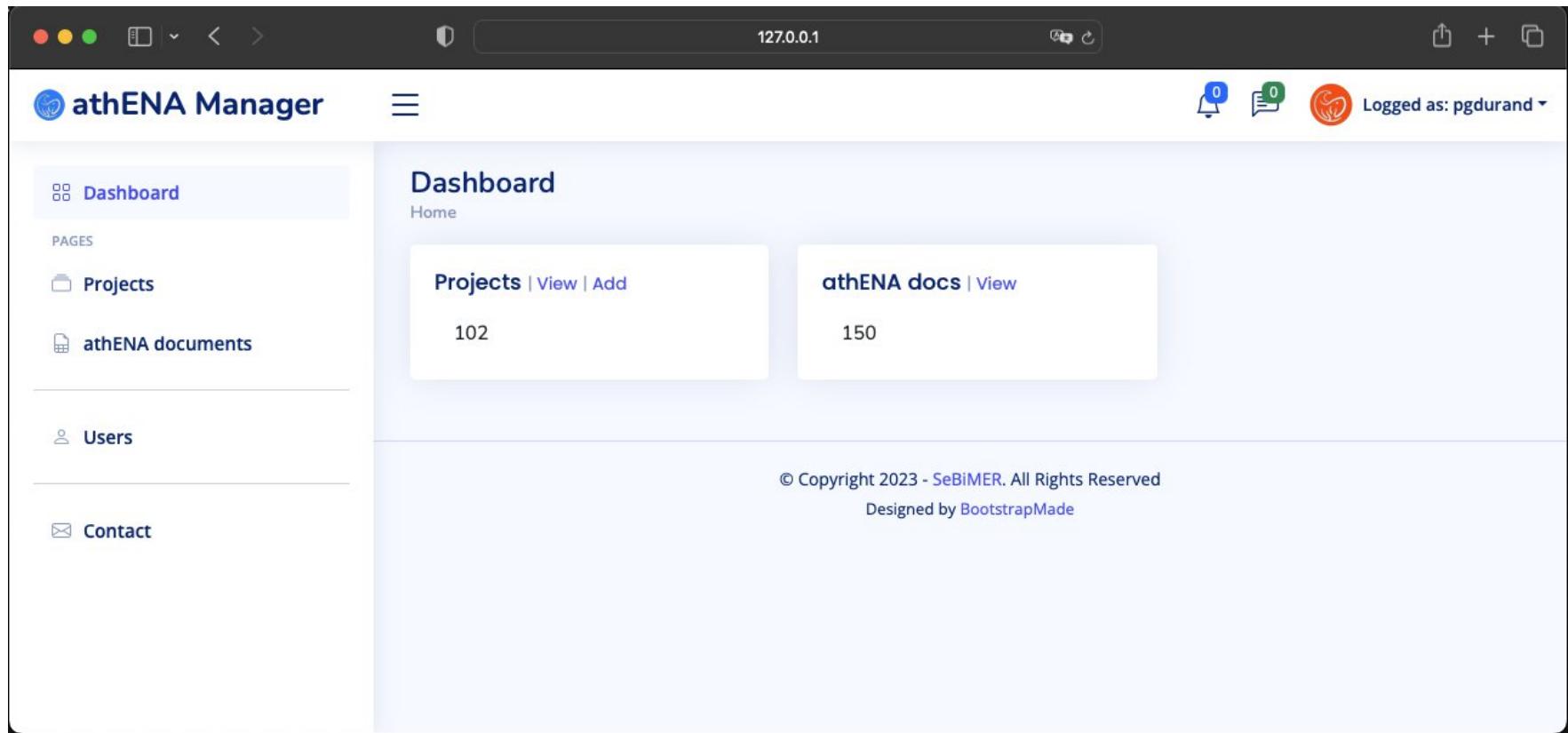


New features (work in progress)

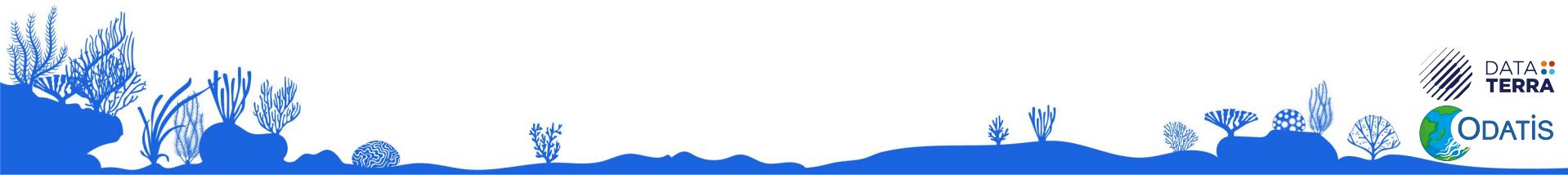
- Programmatic generation of Excel templates
- Template versionning
- Integration of genome submission to main pipeline
- Specific templates (Single-cell)
- Genome annotations



athENA manager (Patrick Durand)



A screenshot of a web browser displaying the athENA Manager dashboard. The URL in the address bar is 127.0.0.1. The page title is "athENA Manager". On the left, a sidebar menu shows "Dashboard" (selected), "Projects", "athENA documents", "Users", and "Contact". The main content area is titled "Dashboard" and "Home". It features two cards: "Projects" (102) and "athENA docs" (150). At the bottom, copyright information reads "© Copyright 2023 - SeBiMER. All Rights Reserved" and "Designed by BootstrapMade". The top right shows a user profile for "Logged as: pgdurand" with notification counts (0) and a log out button.



athENA manager (Patrick Durand)

athENA Manager

Dashboard

PAGES

Projects

athENA documents

Users

Contact

Project: Developpement_larvaire | IFREMER-RBE-RMPF

Home

Information | Update

Creation date: July 26, 2019
Status: PUBLISHED
Manager: Pauline Auffret
Owner: Jeremy LeLuyer
Tech: Patrick Durand

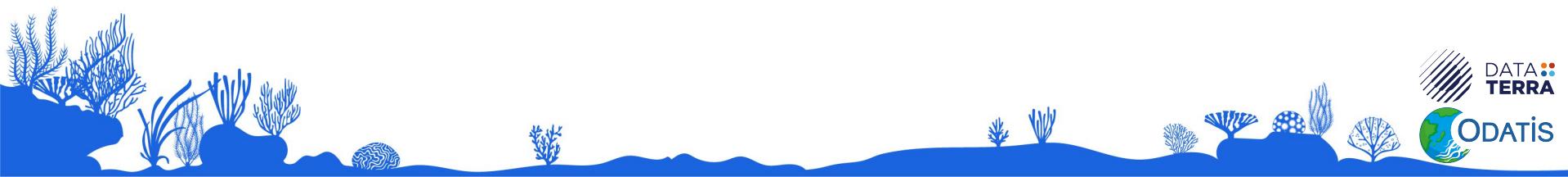
External links | Add link

Action	AddOnLink
	https://sextant.ifremer.fr/Donnees/Catalogue#/metadata/85d63291-71f6-4d36-bfbb-1e0c2a864b7e
	https://www.ebi.ac.uk/ena/browser/view/PRJEB62106
	https://data-dataref.ifremer.fr/bioinfo/ifremer/rmpf/Developpement_larvaire/

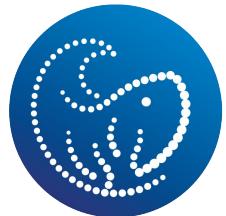
athENA documents | Add athENA

Name	Status	AthENASheetlink	Action
2019_RNA-Seq_Développement-larvaire.xlsx	100	Open athENA document	

Logged as: patrick ▾



Credits



SeBiMER
Bioinformatique marine

SISMER team



Elisabeth Hellec
M2 student + CDD
SeBiMER



Alizée bardon
M2 student + CDD
SeBiMER



Julie Clément
IHPE Perpignan



Bernard de Massy
Institute of Human Genetics,
Montpellier



Frédéric Bigey
INRAe
Montpellier





DATA
TERRA



ODATIS

Thank you !

Pauline Auffret pauline.auffret@ifremer.fr

<https://orcid.org/0009-0001-3834-7670>

<https://gitlab.ifremer.fr/bioinfo/workflows/athena>

To sum-up

