# ERGA IT & Infrastructure
## Committee Meetings

(The last meeting in chronological order must be the first in the document)

| | |
|---|---|
| 🔗 Zoom Link | https://unil.zoom.us/j/99300627815 |
| 🕓 Regular Meetings | Fourth Thursday every other month - 11:00 CET |
| 📁 Committee Drive | ITIC - IT & Infrastructure Com… |
| 📧 Committee Email | itinfra@erga-biodiversity.eu |

📅 ERGA Calendar | 📰 ERGA Newsletter | 🌐 Website

Meeting attendance
https://www.cognitoforms.com/ERGAEuropeanReferenceGenomeAtlas/ERGACommunityMeetingsAttendance

Future meeting topics:

---

## Table of contents

---

# 2026-01-22 Meeting

**Time:** 11:00 - 12:00 am CET
**Join us on Zoom:** https://unil.zoom.us/j/99300627815

**Attendees**, please **register your presence here** | ℹ️ New meeting attendance registration system

**Apologies:** Christian de Guttry (SIB), Rob Waterhouse (CH)

## Agenda

1. Welcome - Tom

2. Review and discuss National Nodes: How To
3. AOB
    a. (Matthieu) CRAM support in ENA
4. Next Meeting: 26th March

## Minutes

1. Welcome - Tom

Welcome to Elena - who will be leading a work package in the BGE+ project. Will aim to develop a solution as to how an infrastructure for biodiversity genomics alongside the iBOL community. Joana is also leading the data WP within BGE+, around improving data processing but also data publishing and provenance. The ITIC could act as a platform for communicating with the community regarding the developments in the project.

2. Review and discuss National Nodes: How To

Expand section on taxonomic updates (updates to sequence name or update to species within biosample)
Elena: What about downstream analysis data and objects, how are these handled? At the moment no guidelines, but data should be linked to a full biosample.

3. AOB

Cram has a new version, which is not supported by ENA. The new version is now the default output by samtools. Joana will follow up internally.
How to submit cram files with embedded information to ENA? Joana will restart the discussions internally.

# 2025-11-27 Meeting

**Time:** 11:00 - 12:00 am CET
**Join us on Zoom:** https://unil.zoom.us/j/99300627815

**Attendees**, please **register your presence here** | ℹ️ New meeting attendance registration system

**Apologies:** Erwan Corre ( ATLASea) (probably late), Christian (participated the first 30 min),

## Agenda

5. Welcome - Tom
6. Review and discuss National Nodes: How To
7. ERGA December Plenary: 2025 ITIC report
8. AOB

## Minutes

4. Welcome - Tom
5. ERGA National Node: How To
    a. Rob mentioned the redundancy with some other ERGA documents, he will highlight them and we can then reduce the text.
    b. Text can perhaps be thought of as more agnostic that just national nodes - but rather “large projects”
    c. Within France: make a sub-branch of Atlasea focusing only those species from Europe to generate the data for the ERGA-FR node and present this to the wider community.
    d.
6. ERGA December Plenary: 2025 ITIC report
    a. Please, add to the slide if necessary including your name if it is not there yet.
    b. Some thoughts for 2026:
    Develop landing pages for all of the portals and sites which collect ERGA data and help communicate the differences between each portal
    Loop into the EBP ITIC communication - focus perhaps more on AI for biodiversity genomics
7. AOB

Partial datasets are uploaded to ENA and can be used by future projects - Is there a way that we can include a credit system in a best-practice document? Document likely to come in the new year.

# 2025-09-25 Meeting

**Time:** 11:00 - 12:00 am CEST
**Join us on Zoom:** https://unil.zoom.us/j/99300627815

**Attendees**, please **register your presence here** | ℹ️ New meeting attendance registration system

**Apologies:** Lada Lukić Bilela (Conference FEBS3+), Erwan Corre ( ATLASea) (probably late, finally here)

## Agenda

1. Welcome - Tom
2. Current working documents:
    a. IT & Informatic Standards Document
    b. National Nodes: How To Guide

Section on “maintaining” a national node required - take experiences from DToL - monitoring and updating existing records

3. Proposal for ITIC-lead ERGA Plenary: November 17th 15:00 CET
4. Update to ITIC Webpage: permanent DOI for ENA guidelines:
   https://doi.org/10.5281/zenodo.14924160
   https://zenodo.org/search?q=parent.id%3A14924160&f=allversions%3Atrue&l=list&p=1&s=10&sort=version
   Could we add the logos/links for infrastructures which ERGA is associated with?
   ENA
   EMBL/EBI-portal
   Ensembl?
   GoaT
   COPO
   Workflowhub?
   Biosamples/bioimage?

   Resources:
   How to get your project in goat https://zenodo.org/records/15826670
   How to make a “community” genome repot https://doi.org/10.5281/zenodo.15634244

5. AOB

Next meeting November 27th
Resources, events and courses from ERGA knowledge hub will now be displayed within TESS

## Minutes

1. Welcome - Tom
2. Current working documents:
    a. IT & Informatic Standards Document

   b. National Nodes: How To Guide
3. Proposal for ITIC-lead ERGA Plenary: November 17th 15:00 CET
4. Update to ITIC Webpage
5. AOB

---

## 2025-06-26 Meeting

**Time:** 11:00 - 12:00 am CEST
**Join us on Zoom:** https://unil.zoom.us/j/99300627815

**Attendees**, please **register your presence here** | ℹ️ New meeting attendance registration system

**Apologies:** Rob Waterhouse (out in Bern all day)

### Agenda

1. Welcome - Tom
2. Current working documents:
   a. IT & Informatic Standards Document
   b. National Nodes: How To Guide
3. Proposal for ITIC-lead ERGA Plenary
4. Invitation to EBP ITIC to join a meeting after the summer
5. Further feedback from Executive Board meeting
6. AOB

### Minutes

6. Welcome - Tom
7. Current working documents:
   a. IT & Informatic Standards Document
   b. National Nodes: How To Guide
      i. CBP portal and brokering: https://dades.biogenoma.cat/
      ii. emilio.righi@crg.eu
      iii. https://portal.atlasea.fr/home/
8. Proposal for ITIC-lead ERGA Plenary
9. Invitation to EBP ITIC to join a meeting after the summer
   a. ERGA used as “test-case” for GoaT in how future projects will be incorporated and follow a given structure.
10. Further feedback from Executive Board meeting
11. AOB
   a. Next meeting end September?

b. ITIC at ERGA Plenary - 17th November, Joana & Alexey, EMBL-EBI and what it can do for you

## 2025-05-22 Meeting

**Time:** 11:00 - 12:00 am CET
**Join us on Zoom:** https://unil.zoom.us/j/99300627815

**Attendees**, please **register your presence here** | ℹ️ New meeting attendance registration system

**Apologies:** Rob Waterhouse (BGE Review), Matthieu Muffatto

## Agenda

1. Welcome - Tom
    a. Summary of the meeting with the EB

Bring in GGBN/GBIF into the meeting to keep up-to-date?

2. AOB
    a. We need clearer, less technical language and step-by-step guides so newcomers can follow ERGA procedures without getting lost.
    b. We could design a survey (modeled on the Goat project) to learn where ENA documentation confuses users and then fix it.
3. How To: ERGA National Node

AtlaSea: Integrated via the committees in ERGA. Discussions ongoing around establishing AtlaSea as an official node within ERGA. How to collect French contributions to AtlaSea under the ERGA umbrella. Other projects sequencing other French organisms (e.g. land) that will not contribute to AtlaSea, but should still be collected. France, Spain, Portugal, Switzerland, and Norway are **"official"** nodes; Germany and Sweden still lack structure. Tom will keep the executive board informed as Portugal and France work out their setups.

From GoaT: How to assign the "primary" project? Is this as easy as taking the "lowest" umbrella associated with each assembly? Issues currently around linking data projects to umbrella projects and vice-versa. It will be blocked at ENA to link to parents (I think this is an issue with ncbi). Assemblies often land in the wrong parent project, so it's difficult to trace who generated what. A revised framework spells out exactly how future nodes must submit and tag data. Cibele will ask NCBI/EBP to remove incorrect project links; Joana will explore a safer linking method so one incorrect link doesn't ripple through ERGA or EBP records.

Key will be creating clear documentation that can be used by nodes of any size and ensure that the end results (data in ENA, tracking updates via GoaT) are consistent and standardised.

4. IT & Informatic Standards
5. **who does what next:**
    a. **Everyone:** review Tom's IT standards + national-node docs and send comments.

   b. **Tom:** stay in touch with the exec board and the Portuguese team about node setup.
   c. **Erwan:** clarify the French node's structure with local leaders.
6. **Next meeting:** Target date is **26 June 2025** (before the summer break) to review survey results, document updates, and node progress.

---

## 2025-03-27 Meeting

**Time:** 11:00 - 12:00 am CET
**Join us on Zoom:** https://unil.zoom.us/j/99300627815

**Attendees**, please **register your presence here** | ℹ️ New meeting attendance registration system

**Apologies:** Rob - I will try to join, not sure I can 🙂, Joana Pauperio, Christian de Guttry

**Minutes were taken by:**

## Agenda

1. Welcome - Tom
2. Rules and guidelines for ERGA National Nodes
   a. Template for BioProject <= is this the latest version?
   b. Guide for creating Umbrella Project

What is their preference? Do they want to administer the project, or by the national node? Should this be coordinated by the counsel reps from each country? Can be delegated.

- [1] Community GOAT sheet (common list) can be used to keep track of which species are being worked on by the various national nodes or smaller projects
  - Access to the sheet needs to be controlled to avoid random people messing it up, the goal is for any ERGA Member from a Country without its own GoaT sheet (currently only Switzerland has their own GoaT sheet) to announce the species they are working on.
  - Access and responsibility to add/update this ERGA-COM GoaT sheet should by default fall on Country Representatives (or they can designate someone from their Country to do this)
- [2] Community BioProject umbrella - similar to GOAT sheet, but here for submitted data and assemblies (currently on Switzerland and France have their own country ERGA BioProject umbrellas)
  - The principle is that any genome assembly can sit under these ERGA umbrellas (even if they don't meet reference quality standards)

    - These should then each be evaluated (EAR, community review certificate) - and their BioProject description should then indicate the link to the EAR as a mark of having been assessed by ERGA Community
- QC stamps such as EAR can be added afterwards as “seal of approval”
    - If they have a lot, then find a better solution
- By country - ERGA community bioproject
    - If more, make a new one for their country

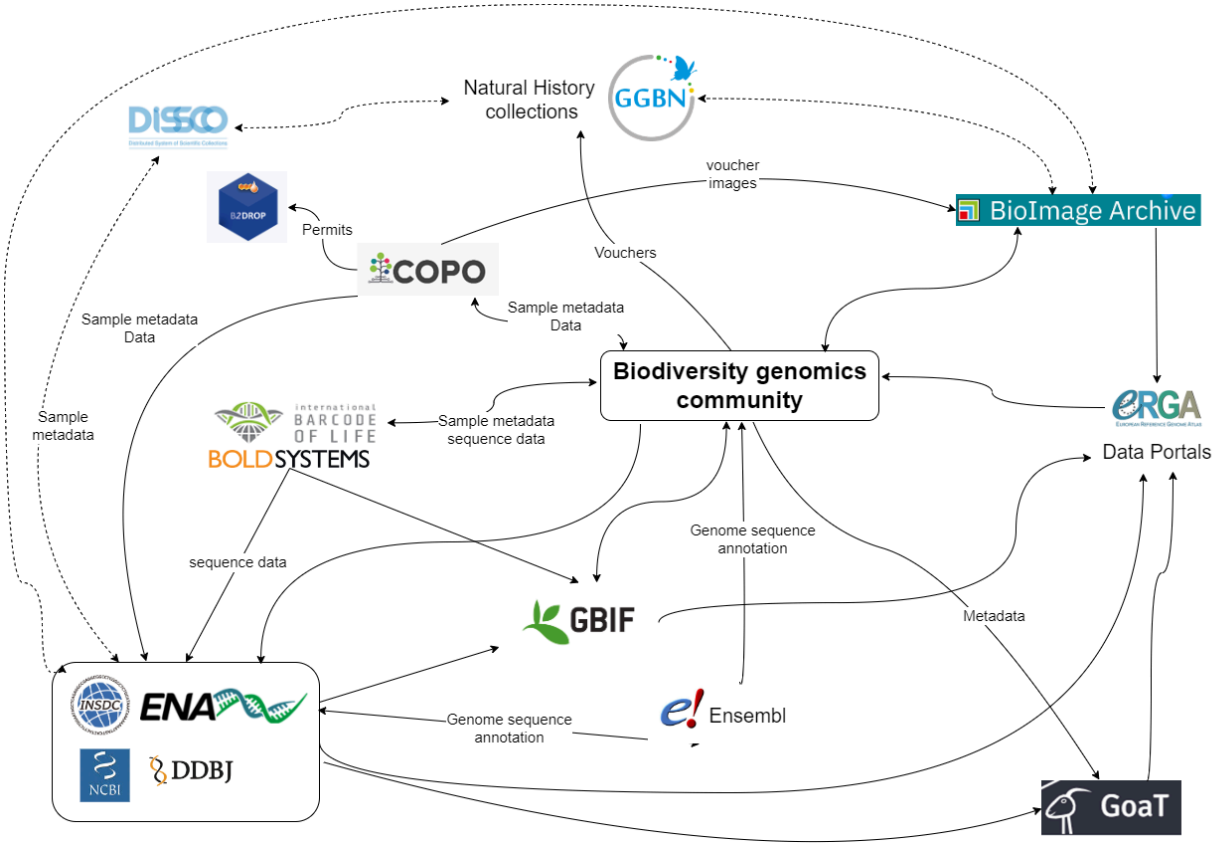
3. Formation of Biodiversity Genomics FDO - Ontology mapping
    a. Current working sheet (Tab 2)
    Ontologies taken from OLS

QC metrics can be listed multiple times, but link to the same URL where appropriate

4. Requirements for inclusion in the ERGA Genome Report Repository
    a. Current draft
        i. Eea map of bio-geographical zones of euro to define ‘what is europe’
        ii. What about waters ?
5. AOB
    a. Start broadcasting guidelines to the wider community for ENA/bioproject

Next meeting 22nd May

New version of the research infrastructure landscape - community centred:
https://drive.google.com/file/d/16OGL3YodmqCmO5f8ryKqFesLrBdKiuL5/view?usp=sharing
(still missing some things, as workflowHub and Ro-Crates - to be added soon)



# 2025-01-23 Meeting

**Time:** 11:00 - 12:00 am CET
**Join us on Zoom:** https://unil.zoom.us/j/99300627815

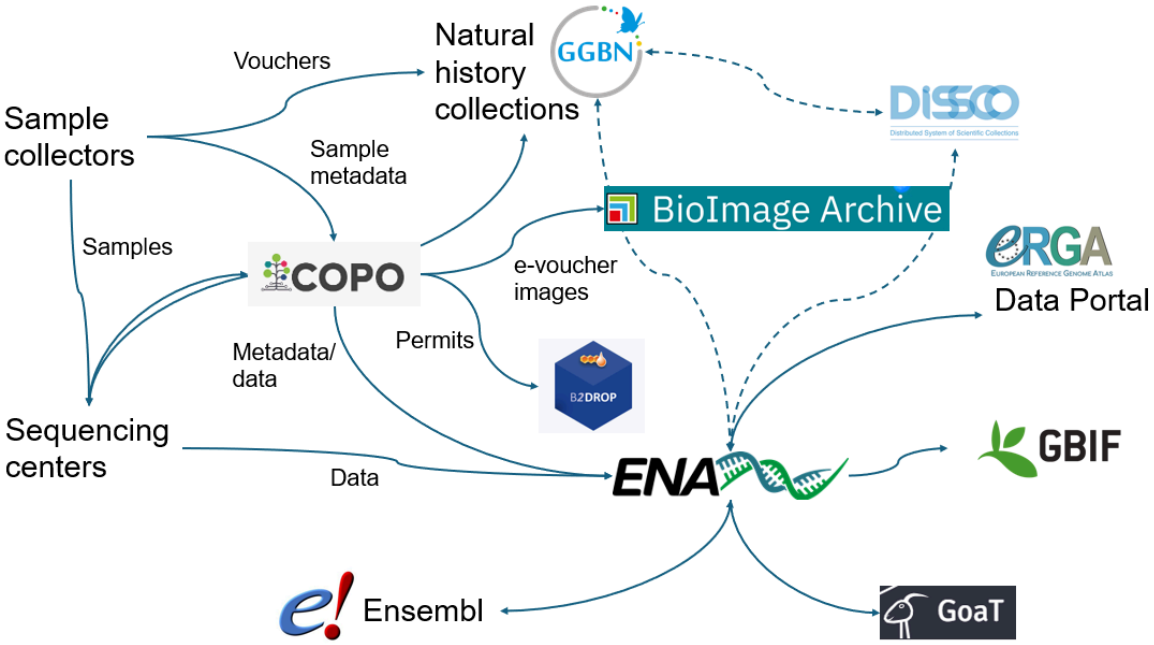
**Attendees**: Tom Brown (IZW), Christian de Guttry (SIB), Jèssica Gómez-Garrido (CNAG), Tyler Alioto (CNAG), Alexey Sokolov (EMBL-EBI), Matthieu Muffato (WTSI, UK), Erwan Corre (FR-ATLASea), Chiara Bortoluzzi (UniFi, SIB), Joana Pauperio (EMBL-EBI)

**Apologies:** Rob Waterhouse (CH), Lada Lukić Bilela

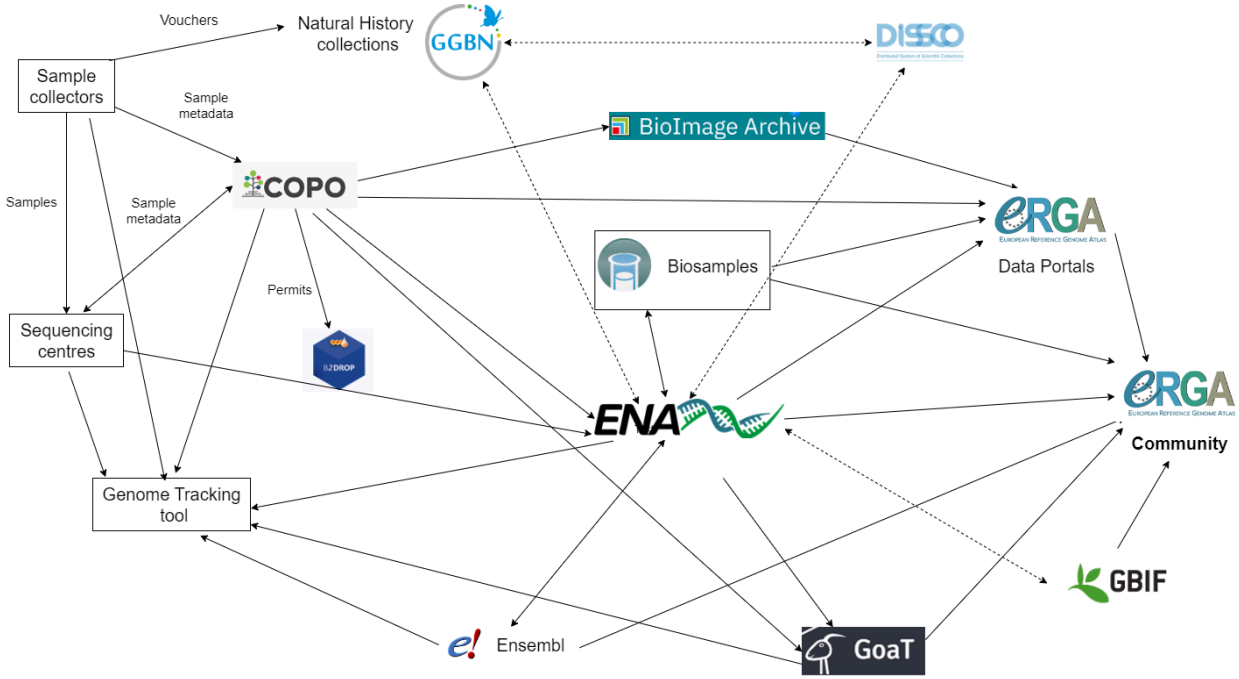
**Minutes were taken by:** AI-assistant

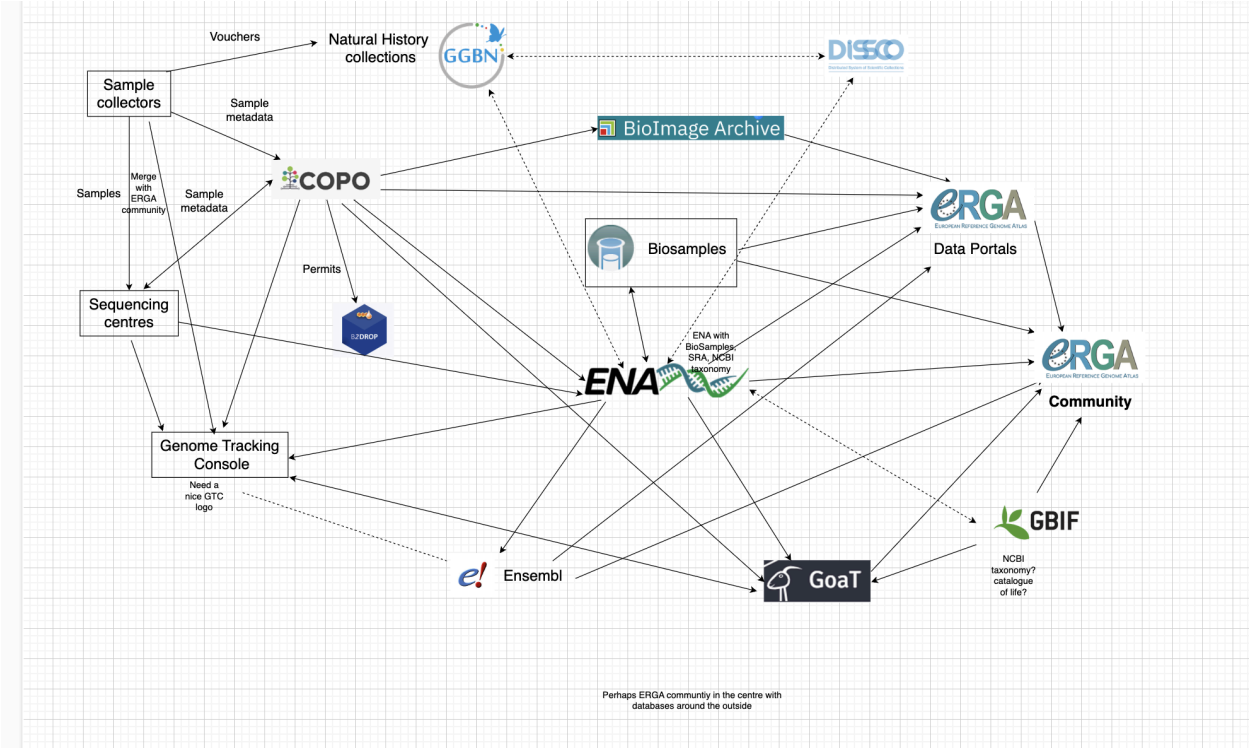
## Agenda

1. Welcome - Tom
2. ERGA ITIC 2025
    a. Goals for better integration of databases




New working version (https://drive.google.com/file/d/1xYb5Eo382znMorOVxdHIqbptHuwSH6E1/view?usp=drive_link):




   b. Better communication of best practices for storing, publishing, and using metadata, data, workflows…
   c. Data management: [Data Management Guide – Biodiversa +]

3. Reference Genome data & metadata linking
   a. *Rupicapra* example

   b. *Culex* example(s)
4. RO-Crates implementation: BGE Case-Studies (Christian) Slides
5. AOB
   a. Proposed ERGA recognition tool: Please review the document and comment if you have any ideas, queries, issues
   b. If time - anyone have experience with the Pensoft ARPHA Writing Tool? ITIC support over the coming year would be great to bit by bit add automation to the creation of tables and figures in the Genome Reports we will publish at Pensoft RIO
   c. We are delighted to announce the EVERSE Network for Research Software Quality Launch Event, which will take place online at 10h CET on Tuesday 18 February. The goal of this network is to improve the policy and practice of research software in Europe and beyond. We welcome everyone interested in software quality to join us for the event. As well as presenting our ideas and plans, there will be ample opportunity for participants to give their input into the goals and activities of the network. Please register for the event at https://indico.cern.ch/e/eversenetworklaunch and let's build a strong and effective Network for Research Software Quality together.
6. Next meeting 27th March @ 11am CET

Minutes

The team discussed the use of AI for meeting minutes, the plans for the Erga ITIC in 2025, and the ongoing improvements in data connection between ERGA and the wider data community.

Next steps

• Joanna to reshape the data flow diagram, putting the ERGA community in the center and merging it with sample collectors and sequencing centers.

• Tom, Joanna, and Alexey to put together a 30-minute presentation about the data flow for an ERGA plenary meeting.

• Tom to provide Joanna with examples of genome sequencing project information for the RO-Crate task.

• IT committee members to review and comment on the draft of the ERGA credit system for community participation.

Summary

AI for Meeting Minutes and Data Integration

Christian proposed using AI to take meeting minutes for efficiency. Tom initiated a discussion about the future of Erga ITIC, focusing on how to better assist Erga members in data integration and metadata thinking. He also highlighted the need for a plan to move forward. Joanna presented an updated map of data metadata links, which was a continuation of her previous work. She also discussed the need for feedback on the direction of the links and the inclusion of the community in the project.

Improving Data Connections and Representation

Joana discussed the ongoing improvements in data connection between Gbif and Erga, aiming to better represent Erga-related data. She also mentioned the need to improve the way data is being pushed and pulled to represent it properly. Christian and Alexey discussed the lack of a direct link between Ena and Ensembl, suggesting that Ensembl should point directly to the Data Portal. Alexey also pointed out the missing links from Ensembl to the Data Portato the Data Portal. Tyler suggested that the connection between Genome Tracking Console and Goat should be bi-directional. Lastly, Matthieu requested a link from GIF to Goat for downloading species occurrence data.

Community-Centric Diagram Restructuring Discussed

The team discussed the restructuring of the diagram, focusing on making it more community-centric. They considered merging sample collectors and sequencing centers with the community, and reducing the number of links. The team also discussed the inclusion of taxonomy and category of life as resources. They agreed to link Ena, Biosamples, and NCBI taxonomy together. The team also discussed the possibility of adding Catalogue of Life as a taxonomic aggregator. The conversation ended with the team deciding to put the community at the center of their project and to link it with sample collectors and sequencing centers.

Reference Genome Creation and Communication

Tom initiated a discussion about the flow of creating a reference genome and the interaction between various committees involved in the process. He questioned whether the information from one committee to another was consistent and whether there was a need for more explicit communication between them. Christian suggested creating teaching modules for the community to better understand the process and best practices. Joana agreed, suggesting that while individuals working on specific aspects of the process may not need to understand the entire process, they should have a general understanding of data management and the importance of the process. Christian shared an example of a data portal issue and suggested that a broader overview of the process could help avoid such issues. Tom agreed that this could be a good task for the committee and planned to follow up with Joana and ALexey about creating a presentation for the Erga community.

Digital Object Creation and Metadata Discussion

Tom discussed the creation of a digital object for a whole genome reference genome, which includes linking together all relevant objects such as sample information, sequencing information, assembly annotation, sequencing data, and the people involved in each task. He mentioned that this object should be primarily machine-readable and secondarily human-readable. Tom also shared a case study of a goat from Slovenia, detailing the process from bio samples to genome assembly and annotation. He questioned whether all biosamples should be included, even if they weren't used for sequencing. Joana and Christian agreed that the level of granularity of the metadata included in these objects should be considered. They also discussed the importance of linking to vouchers and bioimage archives. Diego raised a

question about the contributions section of the report, and Tyler confirmed that not all individuals are listed in the GTC currently, so we still rely on manual intervention.
Addressing Data Entry Inconsistencies
Tyler expressed concerns about the inconsistency in data entry into the tracking console, particularly in sample collection. He mentioned difficulties in matching names and institutions due to variations in spelling and capitalization. Tyler also highlighted the issue of not being able to extract email addresses without logging in as a user. He suggested that requiring an ORCiD for everyone could improve the situation. Tom agreed with Tyler's points and suggested that manual verification and input would be necessary for the small number of case studies they have. He also proposed a hierarchical object system to reduce complexity.
Erga Credit System Development and Launch
Christian discussed the development of a system for credit participation in the Erga community, with Christian seeking feedback from the team. They also discussed the creation of a file with the presence of each committee and their roles, which could be used to justify hours spent on Erga when applying for other jobs. Matthieu announced a new network for research software quality and invited the team to join the launch event. Tom ended the conversation by suggesting the next meeting in late March and encouraging team members to email him with any issues or concerns.

## 2024-12-05 Meeting

**Time:** 11:00 - 12:00 am CET
**Join us on Zoom:** https://unil.zoom.us/j/99300627815

**Attendees**: Alexey Sokolov (EMBL-EBI), Tom Brown (IZW), Rob Waterhouse (CH) Christian de Guttry (SIB), Chiara Bortoluzzi (UniFi), Erwan Corre (ELIXIR-Fr, ATLASea), Joana Pauperio (EMBL-EBI), Matthieu Muffato (WSI)

**Apologies:** Lada Lukić Bilela (UNSA)

**Minutes taken by:** AI Companion

### Agenda

1. Welcome - Tom
2. The ERGA Knowledge Hub - Rob

**How to build a Workflow (Who is the target - early bioinformatician doing their first assembly/annotation/etc?)**
What are your data input and output formats?

What quality control should be performed on your data and outputs? Can these be included as part of your workflow, or should they be considered a separate workflow?
Emphasis on FAIR software - containers, versions, commands, arguments - how are these managed by the various workflow managers?
How much compute is required to process your data? Laptop/HPC/web server?
How are you intending to install your required software? Conda/Singularity/Docker/Include with workflow?
How are you intending to chain the steps of your workflow together?
Snakemake/Nextflow/Galaxy/CWL/Bash/Perl? Pros and cons - learning curves - requirements
What are the necessary steps to document in your workflow README?

- How to structure your data and folders
- How to download and install the workflow and necessary dependencies
- How to run the workflow
- How to view the output
- How to report the software, versions, commands and arguments

Your first workflow

- Show an example workflow doing something not too complicated in NextFlow/Snakemake/CWL(?)
- Can link to Galaxy Training Network for how to create a Galaxy Workflow
- Go through a template workflow step-by-step - how to adapt to your workflow?

How to publish your workflow

- Link to guide - How to submit a workflow to WorkflowHub

Get your workflow reviewed - can other people understand and run it?
Link to other materials - GTN, Elixir FR FAIR Working training

**How to manage your FAIR genomics project**
**How to write a Data Management Plan (Different recommendations for large vs small projects?)**
What should you ask yourself before embarking on your genomics project journey?

- What does my university/institute already provide/mandate regarding data/computing?

How much storage/compute space do you need for your project?

- Can we envision having a simple calculator based on genome size, data types, ploidy, other fields?
- How will you ensure reusability of your data?
- With whom are you sharing data?

What is your short-term (<2 years) plan for data storage?
What is the long-term (10 year) plan for your data storage?

- What are our recommendations? ENA? Tape storage in-house? Cloud storage?

- Ensure data re all linked to each other, which themselves should be long-term databases

How to collect and publish your metadata

- Relates to sample, data, machines, software, versions, licensing, etc

Elixir biodiversity tutorial… coming soon

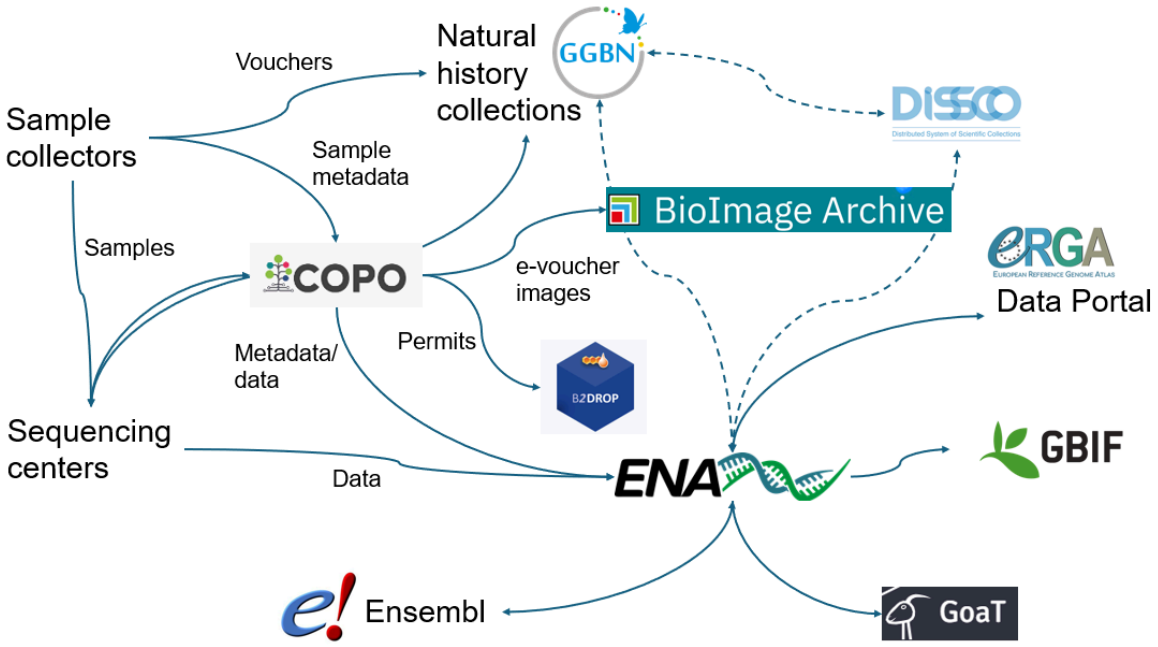
Link to existing tutorials where possible - but how many of these are too broad/detailed/specific?

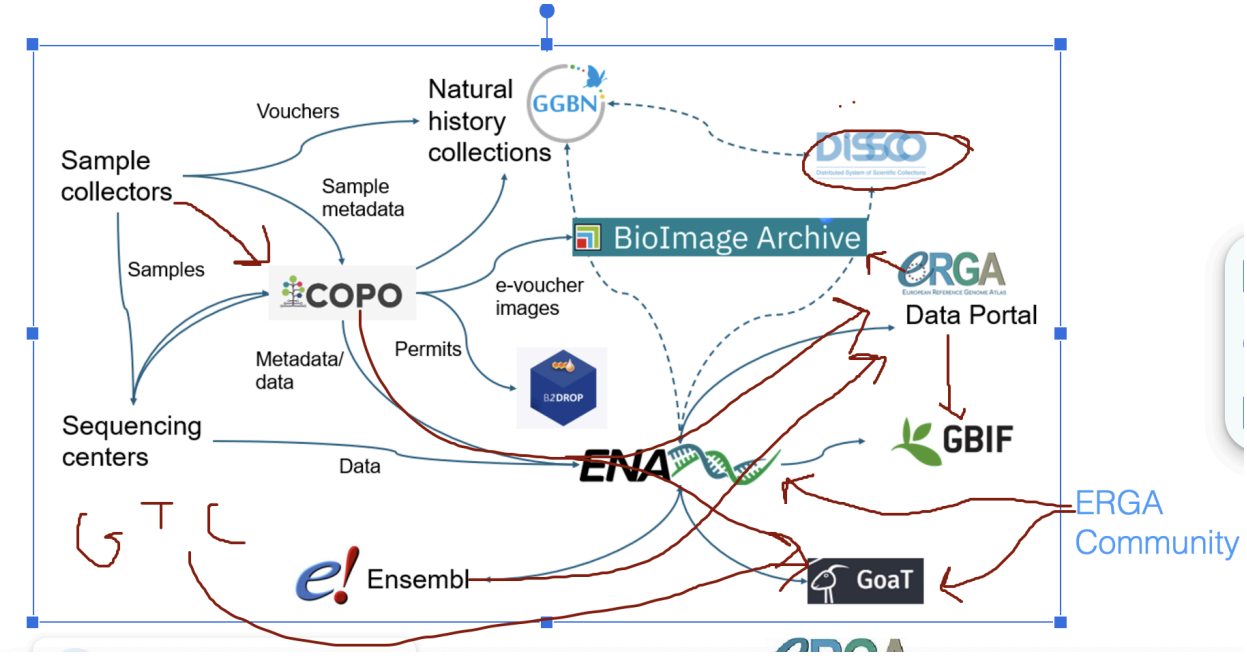
3. AOB

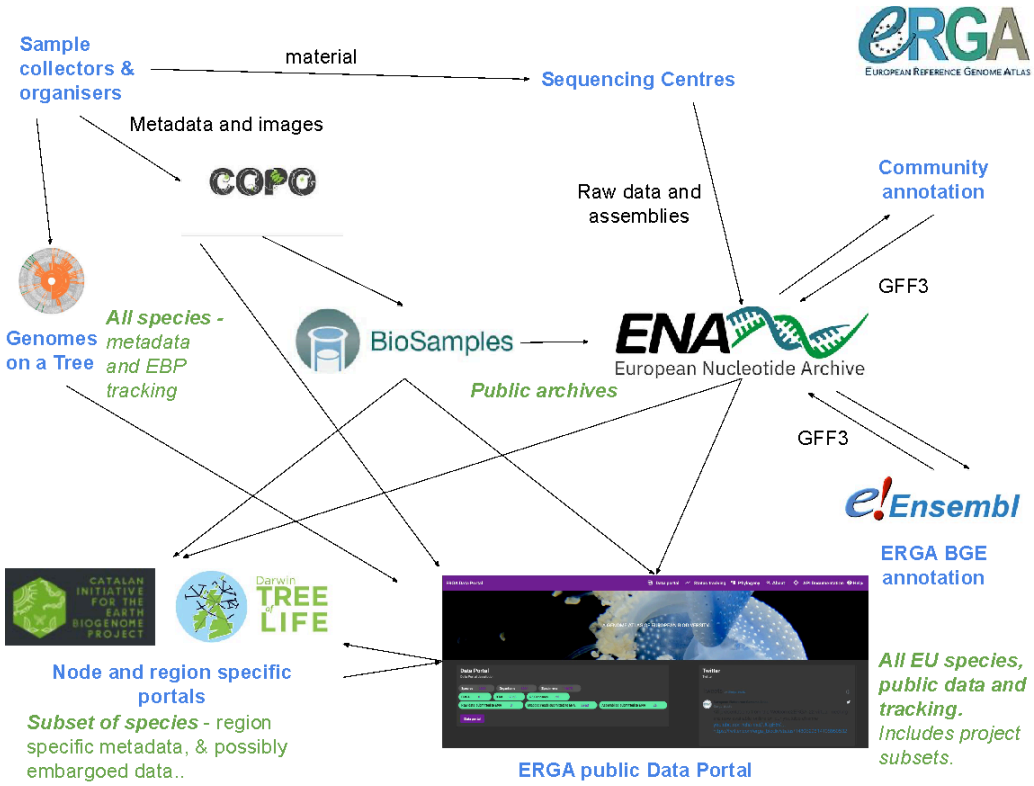
ENA - Umbrellas link to their children - if there are intermediate umbrellas it has been possible to link to parents, but this should change in the future

Bottom-up linking possible in ncbi, complicates matters when mixing ENA and NCBI BioProjects

4. The Biodiversity Genomics [FAIR Research Object] - Tom
5. ERGA Data Infrastructure Map update






6. Next meeting - 23rd January

## Minutes

1. Welcome - Tom
2. The ERGA Knowledge Hub - Rob
3. AOB
4. The Biodiversity Genomics FAIR Research Object - Tom
5. ERGA Data Infrastructure Map update
6. Next meeting - 23rd January

## 2024-10-24 Meeting

**Time:** 11:00 - 12:00 am CEST
**Join us on Zoom:** https://unil.zoom.us/j/99300627815

**Attendees:** Chiara Bortoluzzi (CH), Erwan Corre (FR), Matthieu Muffato (WSI), Rob Waterhouse (CH), Lada Lukić Bilela (BiH), Tyler Alioto (CNAG - joined late)

**Apologies:** Joana Pauperio (clash meeting)
**Minutes taken by:**

### Agenda

7. Welcome - Tom
8. The ERGA Knowledge Hub - Rob
9. AOB
10. The Biodiversity Genomics FAIR Research Object - Tom
11. Next meeting - 28th November/5th December?

### Minutes

7. Welcome - Tom
8. The ERGA Knowledge Hub - Rob

Please familiarise yourselves with the ERGA knowledge hub and the resources that are currently published under the ITIC tag. What is missing? Could we create a set of bite-sized tutorials that would be useful for the community?

9. AOB

How is the pushing of information from BioSamples to GBIF going?

10. The Biodiversity Genomics FAIR Research Object - Tom

Q: which metadata are in COPO but not in ENA ?
MM: If I read our pipeline code correctly, here are the 4 fields that are only in COPO: DESCRIPTION_OF_COLLECTION_METHOD, PRESERVATION_APPROACH, SYMBIONT, SEX

Some from Rupicapra BioSample

DATE_OF_PRESERVATION
TIME_ELAPSED_FROM_COLLECTION_TO_PRESERVATION
PRESERVATION_APPROACH
PRESERVED_BY
PRESERVER_AFFILIATION
BARCODE_PLATE_PRESERVATIVE
COLLECTOR_AFFILIATION
HAZARD_GROUP
SYMBIONT
IDENTIFIED_HOW
SIZE_OF_TISSUE_IN_TUBE
TISSUE_REMOVED_FOR_BARCODING
PURPOSE_OF_SPECIMEN
NAGOYA_PERMITS_REQUIRED
TISSUE_REMOVED_FOR_BIOBANKING
DNA_REMOVED_FOR_BIOBANKING

**Nano-citable object**
*Species Name*
Taxid

Metadata (COPO/BioSamples)
Permits/ethical/legal documents/compliance
Sampling protocol (version/amendments)
Authors related to sampling/identification/preservation
Authors responsible for developing lab protocol

Individuals carrying out the genome assembly
Desired sequencing output (?) *sequencing recipe*
Sequencing instrument/protocol
Accessions of raw sequencing data
Source of sequencing data accession (INSDC/Chinese databank/BOLD/Image archive/clinvar)
Read statistics (Yield/N50 Length/Mean Length/Mean Q-score)

Assembly
    BUSCO score (with version/database/parameters)
    QV
    Contiguity
    Span
Annotation
    Link to gff (structural/functional annotation)

BUSCO on annotation
Other quality metrics?
Workflow/parameters/datasets used
(https://www.rd-alliance.org/groups/fairification-genomic-annotations-wg/work-statement/ reflection)
Link to analysis object (e.g. SNPs)
Link to any subsequent publications

Wrap up, formalise and present to all committees

11. Next meeting - 28th November/5th December?

## 2024-09-26 Meeting

**Time:** 11:00 - 12:00 am CEST
**Join us on Zoom:** https://unil.zoom.us/j/99300627815

**Attendees:** Lada Lukić Bilela (UNSA), Alexey Sokolov (EMBL-EBI), Alex Donath (LIB), Tyler Alioto (CNAG), Matthieu Muffato (WSI), Rob Waterhouse (CH), Tom Brown (IZW), Camila Mazzoni (IZW)

**Apologies:** E. CORRE **(**CNRS - ATLASea program)

**Minutes taken by:**

### Agenda

1. Welcome - Tom
2. COST Action proposal - Chiara
3. Short overview of activities from the BGE over the summer
    a. ENA submission guidelines
    b. ERGA ITIC - Where are we now?
    c. GOAT survey
4. AOB
5. Next meeting - 24th October

### Minutes

1. Welcome - Tom
2. COST Action proposal - Chiara (chiara.bortoluzzi@unifi.it)

18th October deadline for COST action - looking for people to get involved with the action itself later. Focus is on developing best practices in genomic data and its uses. One working group is focussed on best practices in metadata associated with the genomic data. Mostly looking for participants from ITC countries (widening +) as this is a particular focus of the COST.

Focus is on Whole-Genome sequencing data as the standard for genomic research. Focus on the metadata will be on the information associated with the WGS data itself. The idea of the working groups is to develop networks, provide training, without any money for research itself.

3. Short overview of activities from the BGE over the summer
    a. ENA submission guidelines
    b. ERGA ITIC - Where are we now?

DToL is querying in multiple locations - partly out of necessity, partly to ensure consistency. Methods are not currently tracked - have to contact individual teams. Protocols themselves should be in protocols.io for tracking.

    c. GOAT survey

4. AOB

Use of anaconda in academia and research:
https://www.anaconda.com/blog/update-on-anacondas-terms-of-service-for-academia-and-research
Have a look at the ERGA knowledge hub and see if your favourite resource is missing

5. Next meeting - 24th October

## 2024-07-04 Meeting

**Time:** 11:00 - 12:00 am CEST

**Join us on Zoom:**
https://us02web.zoom.us/j/82586151262?pwd=bd8OUx8WsBPascWFNHbT1EGXJREthI.1

Meeting-ID: 825 8615 1262
Kenncode: 702948

**Attendees:** Erwan Corre (CNRS - ATLASea program), Anthony Bretaudeau (CNRS - ATLASea program), Nick Juty (UNIMAN), Joana Pauperio (EMBL-EBI), Matthieu Muffato (WSI), Alexey Sokolov (EMBL-EBI), Christian de Guttry (SIB), Chiara Bortoluzzi (UNIFI - CH), Lada Lukić Bilela (UNSA-BiH)

**Apologies: Rob Waterhouse**

**Minutes taken by:** Tom Brown

## Agenda

6. Welcome - Tom
7. Erwan Corre and the AtlaSea project
   Slides available here
8. BGA24 IT sessions?
9. AOB

## Minutes

1. Welcome - Tom
2. Erwan Corre and the AtlaSea project

BYTE-Sea project providing the digital infrastructure for the AtlaSea project. 8 year program funded by the French funding agency - coordinated by Hugues and Patrick. Aiming to sequence ~4,500 marine species collected from the French coasts (metropolitan and overseas) Call for more projects in the future as genomes are produced focussed on metabolic pathways and environmental studies.

DIVE-Sea sampling teams and expeditions. Biobanking organised alongside the sampling of organisms.
SEQ-Sea Sequencing and assembling of genomes (~4,500 over 8 years). Nature of the organisms, which can also be very small, the project is very ambitious with its targets. Currently ⅔ of samples make it to short-read sequencing, half of those have long-read sequencing and at the moment around 10 have a chromosome-level assembly.

BYTE-Sea coordinating the infrastructure and data management across all of these projects. Focus on integrating each project with public databases such as BioSamples, ENA, WorkflowHub, Goat. Multiple data centers are providing the infrastructure in Paris, Rennes and Plouzane.
Documentation stored in the AtlaSea gitlab https://gitlab.com/peper-atlasea
Developing an AtlaSea Portal, which communicates with the Databases.
https://portal.atlasea.fr/home/ Sample, sequencing, assembly and annotation information should all be present and accessible within the portal.
Create an AtlaSea sampling dashboard so that information can be automatically retrieved to give updates on the progress of sampling.

Development beginning on a dashboard. LIMS is currently managing handling/upload of data, but it is private. First create a dashboard for the AtlaSea consortium to give progress reports and then open it to the public. This will be built on the GTC already developed by CNAG and should be done in close collaboration - likely towards the end of the summer.

BEAURIS framework to create a web-interface which summarises information about the raw data used to generate a genome and assembly including genomic tracks. This should be automatically updated as data are produced and updated.

Functional annotation workflow developed and will be used to generate full annotations for all genomes and these should be uploaded to ENA once finished. Also a task to develop “custom annotation” workflows for a subset of species. These workflows should be public and FAIR following best-practices from the EBP followed by manual curation with Apollo. This is already in place in the BEAURIS framework.

Plan to run comparative genomics automatically on a subset of genomes using existing pipelines.
Also a citizen science project to get students or members of the public to assist in the manual curation of the annotations. Technically it will follow the apollo system with some guidance and automatic validation followed by community efforts to curate gene predictions.

In the field, sampling information will be entered using the CardObs system, which is deployed on mobile devices. This produces a number of google sheets, which must then be collated into the AtlaSea DB and then further integrated into the portal.

3. BGA24 IT sessions?

4. AOB

## 2024-05-23 Meeting

**Time:** 11:00 - 12:00 am CEST

**Join us on Zoom:** https://unil.zoom.us/j/99300627815

**Attendees:** Rob Waterhouse (CH), Christian de Guttry (SIB), Nick Juty (UNIMAN), Joana Pauperio (EMBL-EBI), Lada Lukić Bilela (UNSA-BiH), Matthieu Muffato (WSI), Jèssica Gómez (CNAG), Alexander Donath (LIB)

**Apologies:** Erwan Corre (CNRS-FR)

**Recording:**

**Minutes taken by:** Christian

## Agenda

1. Welcome - Tom
2. AOB
3. iBOL/BGE hackathon feedback - Nick/Joana/Tom
4. Snakemake reporting and RO-crate integration - Tom
5. ENA Projects Guidelines (Rob)

## Minues

1. Welcome - Tom
    a.
2. AOB
    a. We are looking for a new chair and steering committee member. How we can integrate with EBP and ERGA? Vincent is the EBP chair.
    b. There are Biodiversity talks at the Research Data Alliance. The link to recordings will be available during the summer.
    c. World Fair (includes biodiversity use case) : The WorldFAIR Project Webinar Series: Presenting Project Outputs
        i. **TODAY**: https://worldfair-project.eu/event/the-worldfair-project-the-journey-so-far-and-next-steps/
    d. Genome Report: ERGA is producing the Genome Report in a semi-automated way. We are waiting for Sanger to be ready with the fully automated system.
        i. Matthieu: We are waiting for F1000 for the automatization of submission. We talked with people on ENA to attach Genome Note as an object.
    e. For ERGA it is easier if we go together to negotiate with other EBP-affiliated projects.
    f. At SAC we have a task now regarding ‘the rogue genomes’. We would like to gather some statistics from the genome out there. We wanted to take them from Goat. Can we gather these matrics from ENA? We display NCBI metrics that are

standardised. ENA displayed GCA from early last week without the statistics so that the assembly is available for those who need it. Statistics will come later.

3. iBOL/BGE hackathon feedback - Nick/Joana/Tom
    a. At Naturalis (Leiden)
    b. 4 tracks
        i. Data generation and processing: Pipeline and workflows (NHM & Italy) - genome skimming workflows/process
        ii. BOLD release v5: Maily BOLD team.
        iii. Wider integration: we discuss how we can integrate better standards between different databases. Already existing standards. There is the metadata barcoding model now available. The tool will allow us to easily convert from one standard to another. For ERGA this could be good for the barcodes section. From the mapping of standards, you can generate a JSON file.
        iv. The Tree of Life list is more enriched than the barcode metadata model. The issue is mandatory information in the Tree of Life list.
        v. Reference Data Curation: focused on automated curation of BOLD records. At ENA we would like to simplify this list. Given the data volume, we need a more streamlined checklist. We will add a barcoding identifier to sample metadata to have an easier link. The barcode appears as a 'sequence' on ENA.
    c. Tasks will continue and be tracked on the BGE 'alignment' call monthly

4. Snakemake reporting and RO-crate integration - Tom

5. Tom: I was contacted by Vincent to contribute to BGA 2024. There is a list of potential fields.
    a. The EBP IT informatics workshop will be at the conference or the training?
    b. Not clear yet.
    c. We can suggest targeted sessions during BGA, it is more useful for young researchers.
    d. BGA will run over three weeks. They are still gathering speakers.

6. ENA Projects Guidelines (Rob)
    a. This document started as a BGE WP9 (owner=Tyler) output to see how BGE BioProject would sit in the collection/hierarchy of ERGA BioProjects
    b. ENA Project Title and Description
    c. It is now much more of an ERGA document - could ITIC copy this to ERGA GDrive and then work to update it?
        i. Change Satellite to Community

ii. Check that BioProject Titles & Descriptions are correct in this doc, and then that they are updated on ENA
iii. Bring clarity on who is the “owner” of which BioProjects - and make some “rules” about who should do what (children add to parents and/or parents add to children?)
iv. Connect to or integrate the more bio info/technical guidelines that Tom produced How-to guide: ENA
v.

---

## 2024-03-28 Meeting

**Time:** 11:00 - 12:00 am CET

**Join us on Zoom:** https://unil.zoom.us/j/99300627815

**Attendees: Andrea Rezić,** Camila Mazzoni (IZW), Christian de Guttry (SIB), Nick Juty (UK), Matthieu Muffato (WSI), Alexey Sokolov (EMBL-EBI), Sveinung Gundersen (ELIXIR-NO) Beth Yates (WSI), Physilia Chua (ELIXIR Hub), Tom Brown (IZW), Erwan Corre (FR), Alexander Donath (LIB), Martin Pippel (NBIS), Joana Pauperio (EMBL-EBI)

**Apologies: Rob Waterhouse (CH)**

**Recording:**

**Minutes taken by:**

### Agenda

1. Welcome - Tom
2. The European Nucleotide Archive - Joana Paupério
3. ERGA IT & Informatics Standards Document updates
4. AOB

### Minutes

1. Welcome - Tom

   a. A huge thanks to Joana for today's presentation about ENA
2. The European Nucleotide Archive - Joana Paupério
   a. Joana slides.
   b. ENA data and metadata model and developments.
   c. ENA platform for sharing, integration, and dissemination of sequence data, European node if INSDC.
   d. We interact with many partners within EBI and with other data resources.
   e. Reads, Assemblies, and Annotations with their metadata for context about the sample.
   f. A sample checklist of metadata is fundamental. ERGA uses the Tree of Life checklist. Taxonomy is also fundamental.
   g. There are tools for data retrieval.
   h. Working on: large genome submission; Decoupled annotation; Biodata linking at ENA (GBIF, PlutoF, Plazi, GGBN, etc.); Standards for source information (link sequence to voucher information among others); Content enrichment workflows; Data attribution - ORCID claiming;
   i. Question: We're going to submit aligned reads at some point, and I was told it would be as "Analysis objects". Can you tell us where that would be attached on the BioProject hierarchy?
      i. Analysis types are attached to a bioproject. If they derive from read data they can be attached to other objects.
   j. Q: Can we link taxonomic treatment to submission?
      i. Appropriate for cases when a species is described - not necessarily planned within ERGA, but may discover afterwards that the taxonomic classification was incorrect. Treatments could be linked in treatment bank with DOIs.
   k. Could we have a "curation link" to note that an assembly has been curated?
      i. Have an searchable index where individuals can find out if an assembly has been curated which would be user defined.
   l. ERGA could start to add **taxonomic treatment** to our biosamples - to be further discussed
   m. Question: Matthieu asks Joana about adding 'Curated'. She will bring this for internal discussion.
   n. Question: Tom asks about the annotation layer. Work is still ongoing, Joana cannot disclose much at the moment.
   o. Is ENA a permanent stable data repo? Yes. Accessions are stable, connected with INSDC
   p. Clearing house has an API, it is possible to do it programmatically

3. ERGA IT & Informatics Standards Document updates

a.

4. AOB:
    a. Suggestion from Sveinung Gundersen (ELIXIR Norway): Biohackathon Europe project on how to support provenance metadata for genome annotation pipelines, as overlap between "FAIRification of Genomic Annotations WG" in RDA (FAIRtracks continuation), ERGA/BGE, Galaxy/EuroScienceGalaxy and ROCrate/WorkflowHub communities?
        i. Background: Presentation of FAIRification of Genomic Annotations WG to ERGA Annotation Committee, Mar 25 2024.
        ii. Document outline for proposal here
        iii. Want to gather interest from potential co-leads
        iv. Deadline Monday 8 April 2024

## 2024-02-22 Meeting

**Time:** 11:00 - 12:00 am CET

**Join us on zoom:** https://unil.zoom.us/j/99300627815

**Attendees:** Tom Brown (IZW), Alexey Sokolov (EMBL-EBI), Rob Waterhouse (ERGA-CH), Christian de GUttry (SIB),, Nick Juty (UNIMAN), Matthieu Muffato (Wellcome Sanger Institute), Erwan CORRE (ATLASEA program, Roscoff Marine Station), Joana Paupério (EMBL-EBI), Jèssica Gómez-Garrido (CNAG), Diego De Panis (IZW), Paul Davis (WTSI), Maria Angela Diroma (UNIFI), Tyler Alioto (CNAG)
**Apologies:**

**Recording:** Video link

**Minutes taken by:** Christian

### Agenda

1. Welcome - Tom
2. ERGA GTC - Tyler Alioto
3. Development of the ERGA IT & Informatics Standards Document
4. AOB
5. Next meeting 28th March? - Joana Paupério

## Minutes

1. Welcome - Tom
    a. Tyler develop a tool to track the project
    b. If you have question write them in the chat or ask at the end of the call.
2. ERGA GTC - Tyler Alioto
    a. The portal was built under the BGE project
    b. The aim is to collect as much data as possible along the pipeline of high quality reference genome generation. It also serves as a communication tool.
    c. The ERGA data portal comes at the end of the project and it is not focus on communication.
    d. Q&A in the recording
    e. Presentation
3. Development of the ERGA IT & Informatics Standards Document
    a. Go to the document and add comments if you feel like it
    b. This is inspired by EBP but it is more Europe centric and there are some section that have been removed.
    c. It is about best practices, data format, licences, data and metadata structure, protocols, SOPs.
    d. We will collect document and discuss about it during the next meeting.
    e. Are we suggesting or also giving practical suggestions?
        i. The aim is to have a more hands-on document compared to the EBP one
    f. Are we sharing with the other committees?
        i. This is mostly about informatics practices and less about how to collect data
        ii. We could share later with the other committees.
    g. Case-Studies section have being removed.

4. AOB
    a. ITIC to be the gatekeeper to add ERGA Satellite Genomes to ENA Umbrella
5. Next meeting 28th March? - Joana Paupério

---

# 2024-01-25 Meeting

**Time:** 11:00 - 12:00 am CET

**Join us on zoom:** https://unil.zoom.us/j/99300627815

**Attendees:** Tom Brown (IZW), Peter Harrison (EMBL-EBI), Rob Waterhouse (ERGA-CH), Christian de Guttry (SIB), Beth Yates (WTSI), Joana Pauperio (EMBL-EBI), Jèssica Gómez (CNAG), Alexey Sokolov (EMBL-EBI), Lada Lukić Bilela (UNSA_BiH), Nick Juty (UNIMAN), Alice Mouton (UniFI), Tyler Alioto (CNAG), Diego De Panis (IZW)
**Apologies:** Maria Angela Diroma

**Recording:** **ITIC_20240125.mp4**

**Minutes taken by:** Christian de Guttry (SIB)

## Agenda

1. Welcome - Tom
2. EMBL Genome Analysis Team - Peter Harrison
3. Data & Metadata transfer within ERGA & the BGE Project
4. DToL Genome Note generation and integration into the BGE project (Beth Yates)
5. Plans & Visions for 2024
6. AOB
7. Next meeting 22nd February

## Minutes

1. Welcome - Tom
2. EMBL Genome Analysis Team - Peter Harrison Slides
    a. ERGA Public Biodiversity Data Portal. EMBL-EBI realise the potential of big data in biology. We are working with many initiatives worldwide. We closely work with ENA, Ensemble annotation and data portals: Single access point to many data.
    b. We draw data from different archive online and present them in a meaningful way. For ERGA the portal show data and metadata related to the species sequenced. Portals developed on common infrastructure, for economy of scale. Can port features across. Portal displays data in ne place gathered from multiple resources.
    c. We provide several features like the status, when assembly and annotation data are available. Filters are available in the portal to navigate for example the different projects. Also the genome report will be linked when available. The full metadata record could be downloaded. Status tracking is available, the finer

grain tracking is on the BGE portal built by Tyler. The tracking aim is to try help avoid duplicated efforts.

d. Phylogenetic tree based search functionality available
e. It is a multi cloud deployments that uses different technologies i.e. for phylogeny neo4j
f. Data provenance can be tracked to link a genome to a specific project.
g. A challenge is the harmonization across multiple project that differ in their requirements. Nagoya label has been added to concerned species.
h. Next major update will be in February: geographic maps will be available and different dashboard will be available e.g. sequencing technology used.
i. At EMBL we launched the EMBL global portal collecting biodiversity projects. We are continuing to develop it. https://www.ebi.ac.uk/biodiversity/
**j. Questions:**
    i. Camila: Overlap of information between project portal and goat, are you coordinating and discussing with Goat?
    ii. We show data in public domains, we show species when a record is available. Goat show all the species and then assign them to a project. You can see the wishlist from other projects on Goat without a public record.
    iii. ERGA portal is funded by BGE, we can customise it if necessary based on ERGA wishes. Goat present to the global community.
    iv. Tom: Assembly and Annotation quality, how can you retrieve it?
    v. For some species there are difficulties in various steps. We try to find way to present genomes that are available even without labels. Labels have implications. For the quality evaluation about a genome assembly of annotation it will be more subjective. Display is the key.
    vi. Is it possible to reverse-engineer some missing metadata? e.g I put the location, but not coordinates, or I put coordinates but not the samples habitat?
    vii. Missing metadata will be possible to revers engineer but it is out of the scope.
    viii. Publication can be tracked with identifier in the paper. We scan publication and we can connect for example with other projects.

3. Data & Metadata transfer within ERGA & the BGE Project
    a. ERGA is using different infrastructure.
    b. Infrastructure map: complex. document responsibilities. agreements on metadata standards needed. Needs updating. We need to think in term of ERGA and not only BGE.

   c. Review EBP IT and Informatic standards. We need to align the efforts by identifying the area where we are diverging. Tom started a draft of it. Ie ERGA version of https://www.earthbiogenome.org/it-and-informatics-standards-1

   d. Talks about partners contribution are great, we need to keep going.
   e. Identify needs for documentation and training.
   f. We will put in place the scheme fot  highly distributed project and goes partly against the EBP standards. ERGA can bring the model of distributed infrastructure on the contrary of EBP did. ERGA model is tested.
   g. We can facilitate the work of other projects with our example.
   h. Tom will finish the document and share it with the document.
4. DToL Genome Note generation and integration into the BGE project (Beth Yates)
   a. I work on the platform for the automatization of Genome Note generation
   b. The platform was built with the Tree of life but we want it to be available to different projects. A template is created with a defined structure. Pipelines will run on the available data. The portal produces a document that can be sent directly to the publisher.
   c. Data are from ENA or calculated by our pipeline, if info is not available is supplied by third parties.
   d. Current status: We are finalising the ToL genome note and working on the Nextflow pipeline. We are developing an API and UI for the portal.
   e. Our pipeline is written in nf-core Nextflow. All our pipelines are openly available and we have a pipeline website. Genome After Party Data Portal
   f. We can collaborate and use the platform for BGE ERGA
5. Plans & Visions for 2024
6. AOB
   a. https://www.eventbrite.co.uk/e/workflowhub-sops-workshop-tickets-774636708247 (SOPs workshop run by UNIMAN on SOPs metadata)
   b. Joana to present in March
7. Next meeting 22nd February
   a. Tyler will present.

---

## 2023-11-23 Meeting

**Time:** 11:00 - 12:00 am CET

**Join us on zoom:** https://unil.zoom.us/j/99300627815

**Attendees:** Stian Soiland-Reyes (University of Manchester), Rob Waterhouse (SIB/ERGA-CH), Laura, Peter Harrison (EMBL-EBI), Christian de Guttry (SIB), Alexey Sokolov (EMBL-EBI)Paul Davis (WTSI), Tom Brown (IZW), Laura Iacolina (UNISS), Carole Goble (University of Manchester), Beth Yates (WTSI), Jèssica Gómez-Garrido (CNAG), Matthieu Muffato (WTSI) Andrea Rezić (UNIZG-FA), Camila Mazzoni (Leibniz-IZW), Tyler Alioto (CNAG)

**Apologies:** Joana Pauperio (EMBL-EBI)

**Minutes taken by:** Christian and anonymous Shrew (Carole)

**Recording:** Available here

ALL: Please read and modify, if you think that something is inaccurate or not complete!

## Agenda

1. Introduction and Welcome
2. Presentation: Stian Soiland-Reyes & RO-Crate
3. BGE RO-Crate use and development
4. AOB & Next meeting

## Minutes

1. Introduction and Welcome
2. Presentation: Stian Soiland-Reyes & RO-crate
    a. Slides: https://slides.com/soilandreyes/2023-11-23-ro-crate-erga
    b. Aims of FAIR Research Objects. A box that contain all the metadata related to a research object. From WorkflowHub when you download a workflow you will automatically download the metadata related. There is a JSON file describing the content. BY-Covid examples. There is a standard vocabulary ‘Bioschemas’. You can navigate the RoCrate with the menu on the left. There are Ro-crates tutorial on how to create one.
    c. RO-Crate in Biodiversity: wildlife portal to track observation in the wild. Signposting navigate from to landing page to the Ro-Crate. It is a link-API. JSON

file you can see the observation, link to the species, GBIF, description of the images, description of the object.

d. BIODT there is a tutorial on how to generate an RO-Crate. There is a general profile and object specific information.
e. The ERGA manifest is like an RO-Crate. It has already done with DtOL.
f. **RO-Crate is a folder with metadata description.** All the object could be stored in that directory. Detached RO-Crates.https://www.researchobject.org/ro-crate/
g. We have a Trusted RO-Crate for handling the metadata that moves between analysis clients and Trusted Research Environments for federated analysis of patient data https://trefx.uk/5s-crate/
h. RO-Crate is effectively a metadata bag, whereby the metadata specification is defined by a profile. Some contents are files but many are URIs, so there is hashing and checksums to handle that.
i. RO-Crates have a metadata file, and three other kinds of content: files, folders, and URIs.
j. PID resolution is a key part of managing the RO-Crate
k. The *RO-Crate Metadata File Descriptor* contains the metadata that describes the RO-Crate and its content, in particular:
    - Root Data Entity - the RO-Crate `Dataset` itself, a gathering of data
    - Data Entities - the *data* payload, in the form of files and folders
    - Contextual Entities - related things in the world (e.g. people, organizations, places), providing provenance for the data entities and the RO-Crate.
l. FAIR Signposting is using web protocols to resolve to the metadata of web objects
m. RO-Crate Profiles Registry is being developed by our Belgium colleagues
n. Detached and Attached RO-Crate, procedures for handling
o. Hackathon for ENA - openended ness
p. Felix implemented in COPO in under 5 hours
q. The processes of How you use RO-Crates is as important as the RO-Crate itself

3. How could we implement in the ERGA data flow? Who is storing them? Who is defining profiles? Who owns the profile? Where do we store them?
4. We could create a hierarchical structure based on ENA structure. This needs to be maintained.
5. Peter Harrison: how do we do this, who and where
    a. Who owns the profile and who in the process makes and uses them and where do we store them
6. Christian: an opportunity to discuss in the community
    a. Also bring in BioDT
    b. Can we converge

    c. Storage - EUDAT
7. Tom: standardisation
    a. 1:1 with ENA. XML is like
    b. Profiles  Github
    c. ENA requirements are a great seed
    d. Hierarchical RO-Crate
    e. https://app.diagrams.net/#G180Awc6HVXx2ATnx7n5I_62UfrQZPuiVU
8. Tyler: mutability
    a. What is allowed to change
    b. Clear what to do in the Workflows, and point to WorkflowHub URL
9. Peter: Publishing object protocols
    a. When we resubmit
    b. Species names are corrected a lot
    c. Has to go via NCBI
    d. We don't know when it changes in the ENA and then trigger the change
    e. Is the RO-Crate archival or current info (which makes it freshly generated)
    f. Effectively the RO-Crate is split into fixed and snapshot?
10. Peter: generate as we go along
    a. Or do we have it as the final object and its finished and published
    b. Less problems about who generates what and handling different profiles
    c. The profile for  submission will be different to the polished end point
11. We could have intermediate for internal use and final RO-Crates with all the information at the end of a species processing and use it as an archive.
12. TRE-FX the https://trefx.uk/5s-crate/ the review process is included in the RO-Crate
    a. The handles sensitive data, just held by partner institutes
    b. We don't need to be so transparent
    c. Dates?
    d. Its either public
13. Who is accepting RO-Crates today? Journals?
    a. Journals: for workflows, in the EuroScienceGateway - Gigascience, F1000, more in the workflow area
    b. GREIs: DataVerse and InvenioRDM
        i. Mapping of the Metadata from the RO-Crate into the repo
        ii. Using schema mapping to DCAT
        iii. Is the RO-Crate the archive or the active object
    c. Workflows
        i. Update the workflow into WorkflowHub, mints an RO-Crate
        ii. Chain of versioning
        iii. Automated attribution and manual attribution for linking the RO-Crates

d. Other areas are data management plants, HPCs representing reproducible analytics, different lab to move data for data acquisition and storage, EU projects use them a lot, Helmuts association had them in their programme for metadata collection, electronic lab notebook,
e. https://github.com/TheELNConsortium/TheELNFileFormat Electronic Lab Notebooks

14. AOB & Next meeting
    a. Bioscan stream would like to formalise their pipelines. We could maybe help them.
    b. RO-Crates meeting this evening (21:00 CET). 2023 RO-Crate telcons New Zealand colleague on RO-Crates linked to papers.

## 2023-10-26 Meeting

**Time:** 11:00 - 12:00 am CEST

**Join us on zoom:** https://unil.zoom.us/j/99300627815

**Recording available here:**
https://drive.google.com/file/d/14-E4XsKEoZH9f1INgaxitGI-1PzBu7kU/view?usp=sharing

**Attendees:** Christian de Guttry (SIB), Tom Brown (IZW), Rob Waterhouse (ERGA-CH), Tyler Alioto (CNAG), Francisco Câmara(CNAG), Nick Juty (UNIMAN), Erwan Corre (ATLASEA-FR), Alexey Sokolov (EMBL-EBI) Laima Baltrunaite (NRC), Paul Davis (WTSI), Diego De Panis (IZW), Erika Corretto (unibz), Beth Yates (WTSI), Pablo Aguado-Ramsay (UAM-ES), Stian Soiland-Reyes (UNIMAN), Matthieu Muffato (WTSI), Luísa Marins (IZW), Katja Reichel (FU Berlin), Narcis Yousefi (UZH-CH), Iwona Giska (CREAF), Alice Mouton (UniFI), Anna Somogyi (HNHM), Andrea Luchetti (University of Bologna), Rafał Wóycicki (The Plant Breeding and Acclimatization Institute - IHAR, Poland), Carole Goble (UNIMAN), Carlos Fernandes (cE3c, University of Lisbon, ERGA-PT), Jèssica Gómez (CNAG), Roman Volkov (University of Chernivtsy - ChNU)

**Apologies:** Maria Angela Diroma, Physilia Chua, Peter Harrison

**Minutes taken by:** Christian

ALL: Please read and modify, if you think that something is inaccurate or not complete!

## Agenda

1. Introduction and Welcome
2. IT & Infrastructure Committee structure and goals
3. Presentation: Felix Shaw & COPO
4. AOB & November meeting

## Minutes

1. **Introduction and Welcome**
    a. Welcome to everybody. We discuss concepts that exist in ERGA about the infrastructure needed in order to create a reference genome.

2. **IT & Infrastructure Committee structure and goals**
    a. Peter Harrison has been appointed as committee co-chair. He also sits in the EBP IT committee and he is keen to align the efforts of the 2 consortia.
    b. We need to define which are our standard, how we differ from EBP and which are our goals
    c. How are we producing data and the flow of ERGA data.
    d. We are searching for a second Co-Chair. - please email iitinfra@erga-biodiversity.eu if you are interested in getting involved

3. **Presentation: Felix Shaw & COPO**
    a. COPO started 8 years ago. It was a data broker to different repositories including ENA.
    b. Excel was selected as the platform to enter the data, more specifically METADATA. We started with DtOL.
    c. We build SOP to describe metadata of a sample.
    d. We included a validation process. Sanger contributed to it.
    e. User send upload the manifest on COPO, if validated it goes to the sample coordinator of one of the institution. If they also validate the data will be uploaded on ENA and the BioSample created.
    f. We brokered around 30000 samples to date.

g. Example of of the platform works live.
h. Workflow: Add profile: title, description, profile type (you need to request the access to ERGA). Associate profile type: BGE. You can also specify the sequencing centre.
i. In the Action menu you have the list of all the data that you can upload.
j. You can download the blanch manifest and fill it. Once done you can upload it.
k. COPO API explanation.

**4. Q&A**

a. What if a sample is reassigned to a different sequencing facility after COPO submission?
    i. The sample provider need to update at the moment. Sample supervisors could do as well.
b. Does it matter what order you upload to each project? Can I upload annotation before an assembly or sequencing data?
c. Is the Manifest SOP also available together with the blank Manifest (this would be more important than pre-filled lines)?
    i. It could be a good idea. It could come with the black manifest. We will also add the examples
d. Does validation stop once it encounters first error, is it possible to run validation and get summary of all errors?
e. Does COPO check the existence of the permit files?
    i. It matches the filename.
f. How did it link the (fake) permits to the samples ? Through the ETHICS_PERMITS_FILENAME field ?
    i. It matches the filename.
g. Having the Manifest SOP downloadable directly with the blank manifest would really be a great help, as we already know the column names are not really self-explanatory (and samples are so diverse that pre-filled examples may be far off).
h. Who is submitting the reads? For a novice user is great to have a single entry point.
    i. You can have shared profiles. Different people can upload manifest, reads etc.
i. Does the manifest work per sample or per specimen? If I have several tissue samples from one individual, do I put data for all samples under the same specimen_id?
    i. [post-meeting addition: This is all explained in the Manifest SOP.]

j. Can sample providers (and their “helpers”, who do not submit a manifest themselves but perhaps could be “linked”?) also use COPO to track the progress of their samples?

Thanks from all to Felix and his team.

5. AOB & November meeting

---

## 2023-08-24 Meeting

**Time:** 11:00 - 12:00 am CEST

**Join us on zoom:** https://unil.zoom.us/j/99300627815

**Attendees:** Tom Brown (IZW - DE), Christian de Guttry (SIB), Luísa Marins (IZW - DE)**,** Alexey Sokolov (EMBL-EBI), Peter Harrison (EMBL-EBI), Alice Mouton (UniFI), Diego De Panis (IZW, DE), Valentina Galeone (IZW - DE).

**Apologies:** Rob Waterhouse (CH, training); Francisco Pina Martins (FCUL-PT, holliday)

**Minutes taken by:** Christian

ALL: Please read and modify, if you think that something is inaccurate or not complete!

## Agenda

1. Introduction and Welcome
2. Working Document: ERGA Open Data Policy
3. Future meetings schedule
4. Chair and Steering Committee Membership

5. AOB

## Minutes

1. Introduction and Welcome
    a. Keybase invite: https://keybase.io/team/erga.listserv
    b. Most of communications will be done on key base

2. Working Document: ERGA Open Data Policy
    a. We need to have a document about our policy on open data. These recommendations could help members to understand what is needed to become part of EBP and ERGA.
    b. Two documents from EBP are posted at the beginning of the minutes. One is more conceptual and one is more IT.
    c. It reads well and sets out the ERGA position. We need to be more strict about data standards. We cannot only recommend we need to push people to follow our structure otherwise not many people will do this. As the European node we need to follow the data standard of EBP.
    d. EPB does not want to be prescriptive about quality and meeting certain requirements. ERGA can enforce those standards.
    e. We should try to set high standards at least at the beginning. We are in a good position to set standards. We need to guarantee the highest quality. We need to make clear that we follow the highest standard that we have.
    f. This document will go to SSP, ELSI/JEDI and finally to the ERGA Council.
    g. Our goal is thinking about automations of these processes. We will have many genomes at one point and we need to ensure automatization.
    h. People need to understand that it is important to follow standards. It is perceived sometimes as a boring task to retrieve more data but it will be better in the future of scientific research.
    i. The COPO sample manifest is linked to the document. The are almost 50 mandatory field. We maybe need more explicit about how the metadata should be linked to the assembly and annotation data.
    j. Within other project the data are not consistent and we need to make this right. For the people that try to standardise as Alexey is trying to do.
    k. In the CBP there are references used that does not exist anymore. Sometimes different check list are used for the Biosamples. Some of them only have organism name.
    l. COPO should be strongly suggested and sold as an easier solution. In the future maybe more projects will follow this (Hopefully).

m. Also ENA is trying of standardise the metadata. We cannot have manual curation, we need to automatise the process. The standards to be follow are the mandatory of COPO.
n. We don't need to replicate everything in this document. We need to guide people to use COPO and link to those resources so we will not have to update our policy often.
o. Data type: We need to be more comprehensive. We could link this to ENA because those are the data that they accepted. For the sequencing data and annotation we can reference ENA.
p. For annotation GFF3 are waiting to be accepted as the official format. We don't need more formats.
q. Automated Genome Note: How do we retrieve metadata and data. In BGE it is as soon as possible and we could also suggest/enforce/strongly recommend the same in this document.
r. Softwares: We recommend WorkflowHub or github. Ask COPO if their scripts are public and on which platform.
s. Ethics: How can we ask the permits to the researchers? Talk with Amber about liability. ERGA is not liable. ERGA can take measures if needed.
t. Specify that permitting is related to European researchers and species.

3. The document will be sent to the ITIC committee again this week and hopefully in 2 weeks to the SSP committee.

Current working document

4. Future meetings schedule

2023-09-28 Felix?
2023-10-26 Stian?
2023-11-23 Alexey
2023-12-28 - No meeting
2024-01-25 Peter & Alexey
2024-02-22

5. Chair and Steering Committee Membership

Definition of roles

6. AOB

# 2023-07-27 Meeting

**Time:** 11:00 am CEST

**Join us on zoom:** https://unil.zoom.us/j/94341750871

**Attendees:** Tom Brown (IZW), Peter Harrison (EMBL-EBI), Alexey Sokolov (EMBL-EBI), Jèssica Gómez (CNAG), Tyler Alioto (CNAG), Maria Angela Diroma (UNIFI)

**Apologies:** Rob Waterhouse (CH, travelling back from SMBE), Francisco Pina Martins (FCUL, supervising exam), Diego De Panis

**Minutes taken by:** Tom Brown

ALL: Please read and modify, if you think that something is inaccurate or not complete!

## Agenda

6. Introduction and Welcome
7. Use of Keybase and Google Groups
8. Explanation of Committee Roles and formation thereof
9. Volunteering to present at future meetings
10. AOB

## Minutes

7. Introduction and Welcome
8. Use of Keybase and Google Groups

Please joining the #Committee_ITinfra channel on the ERGA Keybase for future updates

keybase://team-page/erga.listserv
https://keybase.io/team/erga.listserv

I will create a google group after the meeting, to allow people to freely join/leave as they wish

9. Explanation of Committee Roles and formation thereof

Chairs
Steering Committee
Coordinator

Peter - How well have we captured the IT&Infra work going on across ERGA/BGE. Have we captured all work packages?

10. Volunteering to present at future meetings

**Currently approached:**

Felix Shaw - Overview of COPO for data and metadata brokering (August?)
Stian Soiland-Reyes - Packaging research artefacts with RO-Crate.
Peter Harrison/Alexy - ERGA Data Portal (January)
Tyler Alioto - ERGA Tracking tool (September)

**Future Meetings Wishlist:**

FAIR publishing of pipelines & workflows
Use of public resources for compute (e.g. Galaxy/AWS)
Setting up a local Galaxy Instance (UniFI)
Protocols.io space setup by Scilife to be used by ERGA/BGE (WP4/5/6/7/11 to use this space) - Olga to run a training on how to use protocols.io (mid-Sep)
Workflowhub for publishing workflows
Definition of key terms and full text for BioProjects

11. AOB

List of Members and roles - what are the individual goals within ERGA/BGE/EBP?
EBP standards driven by VGP/DToL - not much further discussion of data/metadata standards
Creation of Umbrella BioProjects for ENA
https://docs.google.com/document/d/16G5BvKf31T_IUhB8Wbwd7bMmrTursdZSVERpBSRZbWI/edit?usp=sharing

# 2023-06-29 Meeting

**Time:** 11:00am CEST

**Join us on zoom:** https://unil.zoom.us/j/94341750871

**Recording:**
https://unil.zoom.us/rec/share/rjFsVm8gAGFhUlh061SyjEe-B3t1SrtlE98hVwBupYsGOe3_ftiDSgxYauGJAmE.Ft7PJgCP1tn-LWsU
**Passcode:** qg&9$ul1

**Attendees:** Please write your name (affiliation)
Christian de Guttry (UNIL), João Pimenta (CIBIO-BIOPOLIS), Camila Mazzoni (IZW, DE) Alexey Sokolov(EMBL-EBI), Tom Brown (IZW, DE), Joana Pauperio (EMBL-EBI), Peter Harrison (EMBL-EBI), Matthieu Muffato (Wellcome Sanger Institute), Sharif Islam (Naturals/DiSSCo), Francisco Pina-Martins (cE3c/University of Lisbon), Nick Juty (UNIMAN) ,Alexander Donath (Leibniz LIB) ,Yutang Chen (Molecular Plant Breeding group, ETH, Zurich) ,Maria Angela Diroma (UNIFI), Robert Waterhouse (UNIL), Jèssica Gómez (CNAG), Tyler Alioto (CNAG).

**Apologies:** Alexandru Mizeranschi

**Minutes taken by:** Christian

ALL: Please read and modify, if you think that something is inaccurate or not complete!

## Agenda

1. Introduction and Welcome
2. Current tasks being undertaken under the “ITIC” umbrella
3. Establishing Committee goals and structure
4. AOB

## Minutes

1. **Introduction and Welcome**
    a. Alex Donath: Leibniz LIB head of computational genomics and hpc
    b. Francisco Pina-Martins: Bioinformatics at cE3c/University of Lisbon. Very interested in reproducible workflows.
    c. Felix Shaw: From COPO and I'm a data scientist/software engineer
    d. Peter: Genome Analysis team leader EMBL-EBI. My team is developing the ERGA data portal
    e. Alexey: EMBL-EBI Member of Peters team and project lead for the ERGA data portal
    f. Joana: EMBL-EBI more specifically ENA and data and metadata management and standards for biodiversity projects, including ERGA/BGE
    g. Mariangela: University of Florence and I deal with data management.
    h. Camila: Leibniz institute IZW. Chair of ERGA and I wil coordinate the BGE genome team until 2026
    i. Joao: I work at Cibio leading 2 tasks of BGE project and coordinator of DAC
    j. Matthieu: Tree of Life (DToL and others) at Sanger. We develop pipelines for genome assembly and curation, and genome analysis. We also do data management and curation, incl. GoaT.
    k. Sharif: data architect at Naturalis Leiden and I cover some tasks in BGE specially the FAIR data.
    l. Nick: Manchester University BGE WP12 and WP9. We work on RO-Crate
    m. Tyler: CNAG in Barcelona and I'm in charge of assembly delivery. I'm also involve in submission and tracking of data.
    n. Jessica: CNAG in Barcelona I develop pipeline for assembly and annotation.
    o. Rob: UNIL and SIB, vicechairs of ERGA and in charge of WP9 and WP12
    p. Yutang: ETH in Zurich wrokin in the molecular lab. PhD assembly and phase very large heterozygous genome.

2. **Current tasks being undertaken under the "ITIC" umbrella**
    a. Tom presentation Slides.
        i. EUDAT for storing legal data from sampling to publishing.
        ii. Protocols.io is the platform that will host our SOP and protocols.
        iii. The manifest is hosted in the ERGA Github.
        iv. We are creating an ERGA standard set for assembly and annotation
        v. Peter and Alexey are responsible for the ERGA portal.
        vi. Tyler produced the BGE-ERGA tracking system.
        vii. Pipelines will be stored also on Workflowhub.
        viii. Downstream analysis: Are we able to create standardize pipeline for the many research branches?

ix. ERGA website will become the landing site for all the questions. We have a Wix account.

b. DtOL users comments: pipelines are needed with more details.
c. Related to what you said, we are building what we call a “Universal Genome Note” platform. The goal is that anyone could enter their assembly details, and the service would generate a genome note according to a template of their choice, and facilitate submission to the journal.
d. We can cite the DOI of the genome note and the version of the pipeline used
e. If there are citable workflows is ok. Modifications should also be listed, even the manual modifications.
f. Workflow hub could be use to cite pipelines and all the modifications to the pipeline. The workflow is citable.
g. The software used for a specific assembly should be specified. We can think about this.
h. Also the config files should be available.
i. We need for reproducibility all the file connected to a workflow.
j. Tomorrow there is a workload

**3. Establishing Committee goals and structure**

a. We would like to formalize the structure:
   i. Chair, steering committee, sub committees.
   ii. We need to define what are our goals in ERGA, BGE and EBP as a whole.
   iii. Frequency of the meetings.
   iv. Presentation of the expertise of people
   v. We can maybe start to identify broad topics. Metadata, reproducibility, etc. Once we have those topics everything will be smoother.
   vi. DtOL is developing a genome portal to generate Genome Notes. We are asking for help from other consortia/institution.
   vii. It would be great to also add the sampling stage in the genome notes/reports.

4. AOB

Ideas for future meetings:

- Finer details of citing pipelines and workflows? Can we include pipelines with options (e.g. assembler A or B, scaffolder C or D) or does each individual workflow require a different workflow and doi?