# Elixir biohackathon proposal

Submission form at
https://docs.google.com/forms/d/e/1FAIpQLScoid8e3N5-i-RJNQoFblv0gB1XJfeVlMwVrNOMG3V991nKYw/viewform

# Data clearinghouse, validation and curation of BioSamples/ENA/Breeding API endpoints/MAR databases

## Name

Data clearinghouse, validation and curation of BioSamples/ENA/Breeding API endpoints/MAR databases

## Leads


- Luca Cherubin <cherubin@ebi.ac.uk>
- Melanie Courtot <mcourtot@ebi.ac.uk>
- Philippe Rocca Serra <proccaserra@gmail.com>
- Nils Peder Willassen <nils-peder.willassen@uit.no>
- Cyril Pommier <cyril.pommier@inra.fr>


## Background (712 char, max 750 characters)

Regular practice in life science is to curate and publish high quality metadata describing experiments performed on biological sample(s).
Unfortunately this curated, high quality metadata is often not linked to the original assay and results, reducing data FAIRness.

A possible solution to this issue would be to integrate this diversity of resources by
1. Having metadata that can be easily extracted with bots/crawlers,
2. Storing it in a central repository and

3. Validating it against predefined schema.

These overlap with the goals of different ELIXIR implementation studies, including Bioschemas, Data Validation and the Establishment of an ELIXIR Contextual Data Clearinghouse Implementation study.

## Expected outcomes (491 char, max 500 characters)

A small subset of publications and curated data resources will be selected to build a proof of concept of an accession mapping - data clearing house - data validation and curation service to support the positive feedback loop to BioSamples, ENA and BrAPi repositories.
We will:

- Create an accessioning mapping service which leverages literature references to establish correspondence between services
- Develop a clearinghouse for curation storage
- Define JSON schema(s) and validate the metadata

## Expected audience

- Developers interested in Bioschemas applications
- Developers with knowledge on any of JavaScript, Java, GO, Python, data indexing tools
- Data resource developers and owners
- Curator and data validators
- Ontologists

## Number of expected hacking days

4

## Estimated number of participants

**5/10**

## Related works and references

- Human Cell Atlas metadata schema validation - https://github.com/HumanCellAtlas/ingest-validator

- EMBL-EBI Unified submission interface - https://github.com/EMBL-EBI-SUBS/json-schema-validator
- Elixir data validation implementation study - https://www.elixir-europe.org/about-us/implementation-studies/data-validation-2018
- Elixir bioschemas http://bioschemas.org/
- MarRef bioschemas extract demo https://github.com/EBIBioSamples/bioschemas_marref_demo
- HipSci publication https://www.ebi.ac.uk/biostudies/studies/S-BSST16
- MarRef database https://mmp.sfb.uit.no/databases/marref/#/
- Dataflow for the clearinghouse https://docs.google.com/drawings/d/1olEoapldmrJDfTRomHeCFdpxhJknJ7GFJO-WEMYzp5Y/edit
- Establishment of an ELIXIR Contextual Data Clearinghouse https://www.elixir-europe.org/news/new-portfolio-implementation-studies-selected-data-platform

## GitHub or any other public repositories of your FOSS products (if any)

- https://github.com/HumanCellAtlas/ingest-validator
- https://github.com/EMBL-EBI-SUBS/json-schema-validator
- https://github.com/EBIBioSamples/bioschemas_marref_demo
- https://github.com/BioSchemas/specifications