ENA: Improving Experiment Metadata Standards

focusing on sequencing experiment type checklists

Peter Woollard + Josie Burgin & Guy Cochrane

Data Standards Biocurator, Data Coordination and Archiving

European Nucleotide Archive, EMBL-EBI, UK

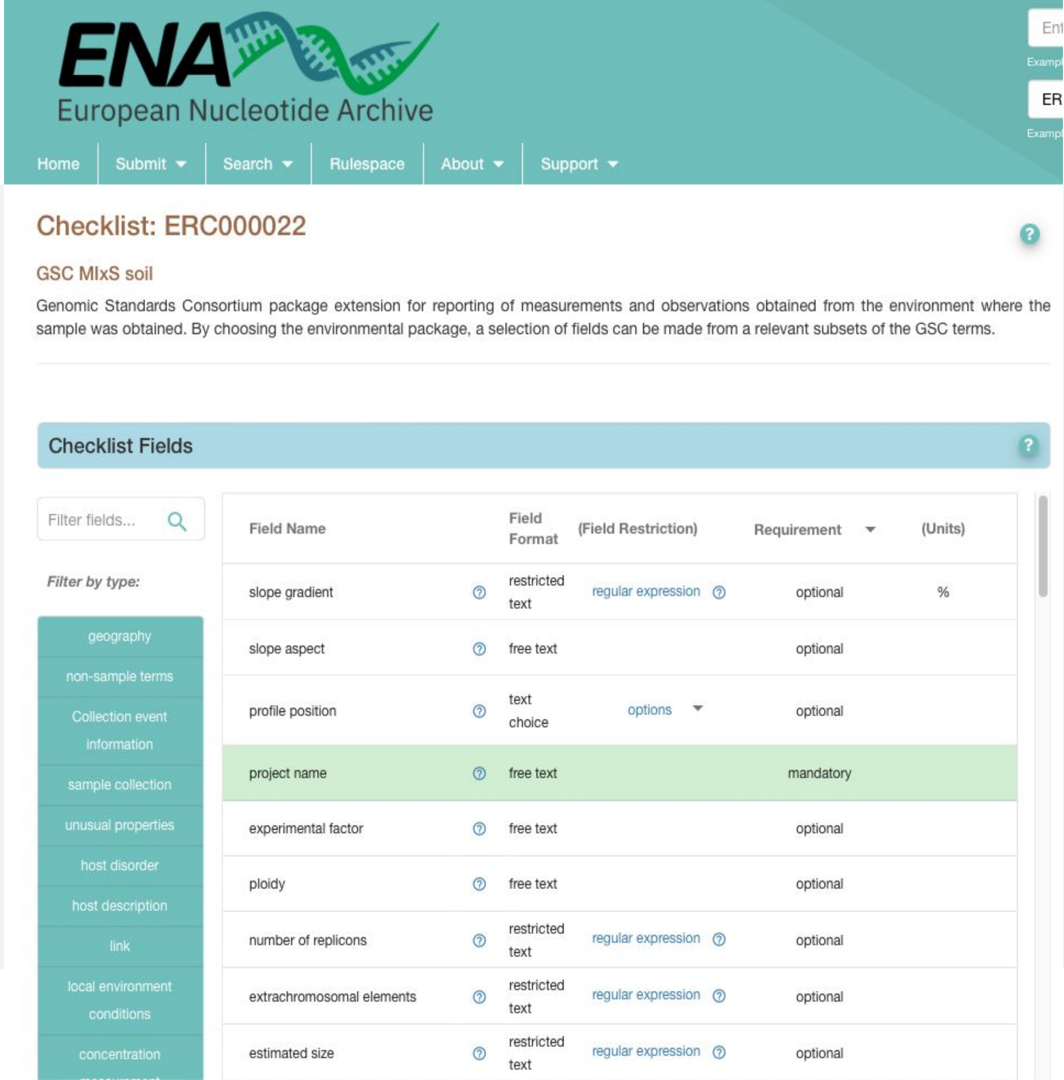
What are ENA and the International Nucleotide Sequence Database Collaboration (INSDC)?

- INSDC:
  - A long-standing foundational initiative that operates between DDBJ, EMBL-EBI and NCBI https://www.insdc.org/
  - Covers the spectrum of data raw reads, through alignments and assemblies to functional annotation, enriched with contextual information relating to **samples** and **experimental** configuration
  - All sequence data and contextual information(metadata) are shared
- European Nucleotide Archive (ENA):
  - The part of EMBL-EBI focused on nucleotide sequencing information
  - Recognised as a Global Core Biodata resource

GLOBAL BIODATA COALITION

Checklist

- A checklist is a set of fields and values
- The purpose is to collect consistent metadata collection
- Fields may be mandatory or optional
- Values may be controlled or free text
-
- A “sample” checklist example is on the right



Checklist: ERC000022

GSC MIxS soil

Genomic Standards Consortium package extension for reporting of measurements and observations obtained from the environment where the sample was obtained. By choosing the environmental package, a selection of fields can be made from a relevant subsets of the GSC terms.

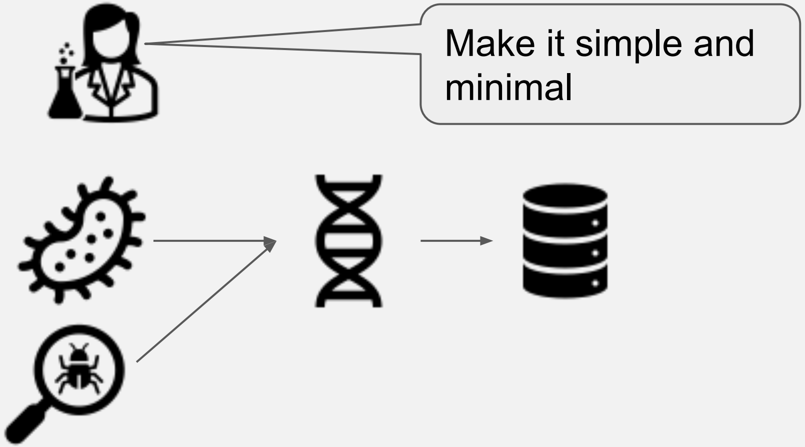
Checklist Fields

Field Name	Field Format	(Field Restriction)	Requirement	(Units)
slope gradient	restricted text	regular expression	optional	%
slope aspect	free text		optional	
profile position	text choice	options	optional	
project name	free text		mandatory	
experimental factor	free text		optional	
ploidy	free text		optional	
number of replicons	restricted text	regular expression	optional	
extrachromosomal elements	restricted text	regular expression	optional	
estimated size	restricted text	regular expression	optional	

Why do we need a Sequence Experiment Checklist System?

Two main use cases:

1) Generate and Deposit data

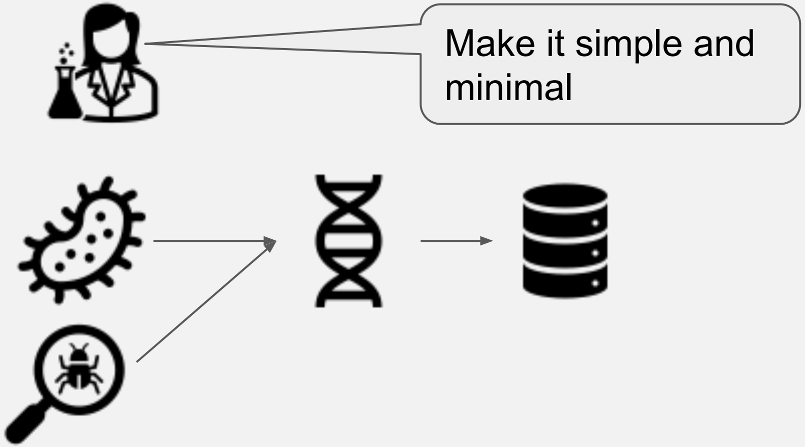


Make it simple and minimal

Why do we need a Sequence Experiment Checklist System?

Two main use cases:

1) Generate and Deposit data



Make it simple and minimal

Increasingly Complex: NGS etc., analysing much of the cell and environment

metabarcoding, chromosome structure, transcriptomics etc

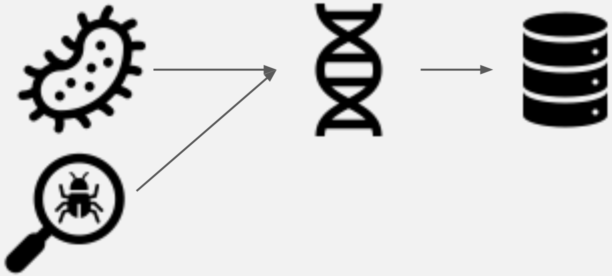
Many dependencies between library_selection, library_strategy etc.

Why do we need a Sequence Experiment Checklist System?

Two main use cases:

1) Generate and Deposit data

Make it simple and minimal



Increasingly Complex: NGS etc., analysing much of the cell and environment

metabarcoding, chromosome structure, transcriptomics etc

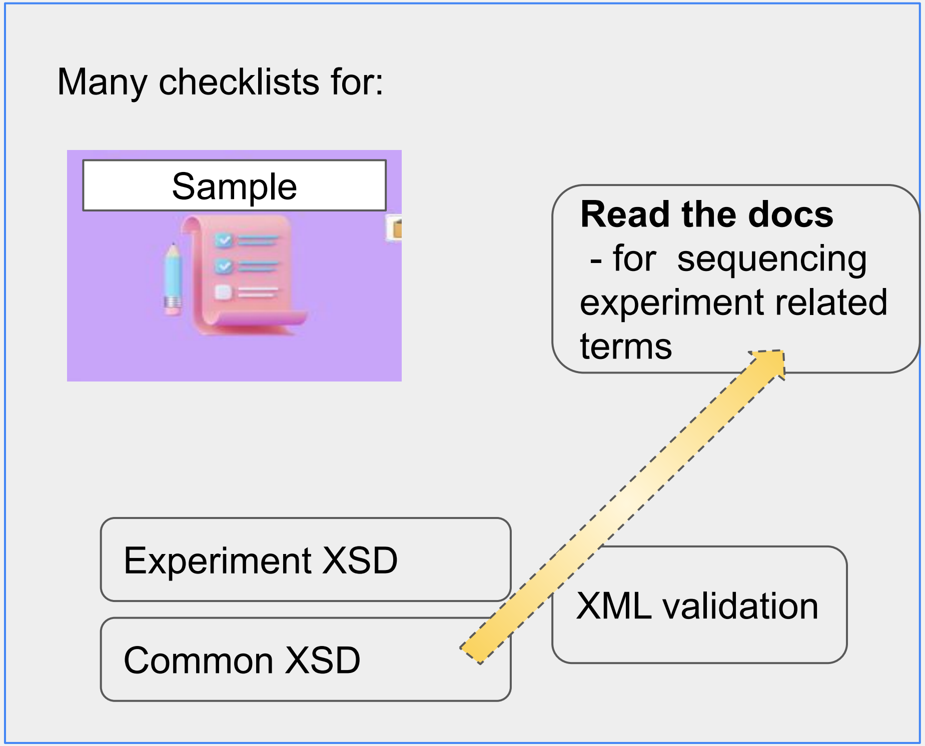
2) Find and re-use the ever increasing data

All easier, if data follows: https://www.go-fair.org/fair-principles/

N.B. Investigational experiment metadata in biosamples

Many dependencies between library_selection, library_strategy etc.

What Exists in ENA?



Many checklists for:

Read the docs
- for sequencing experiment related terms

Experiment XSD

Common XSD

XML validation

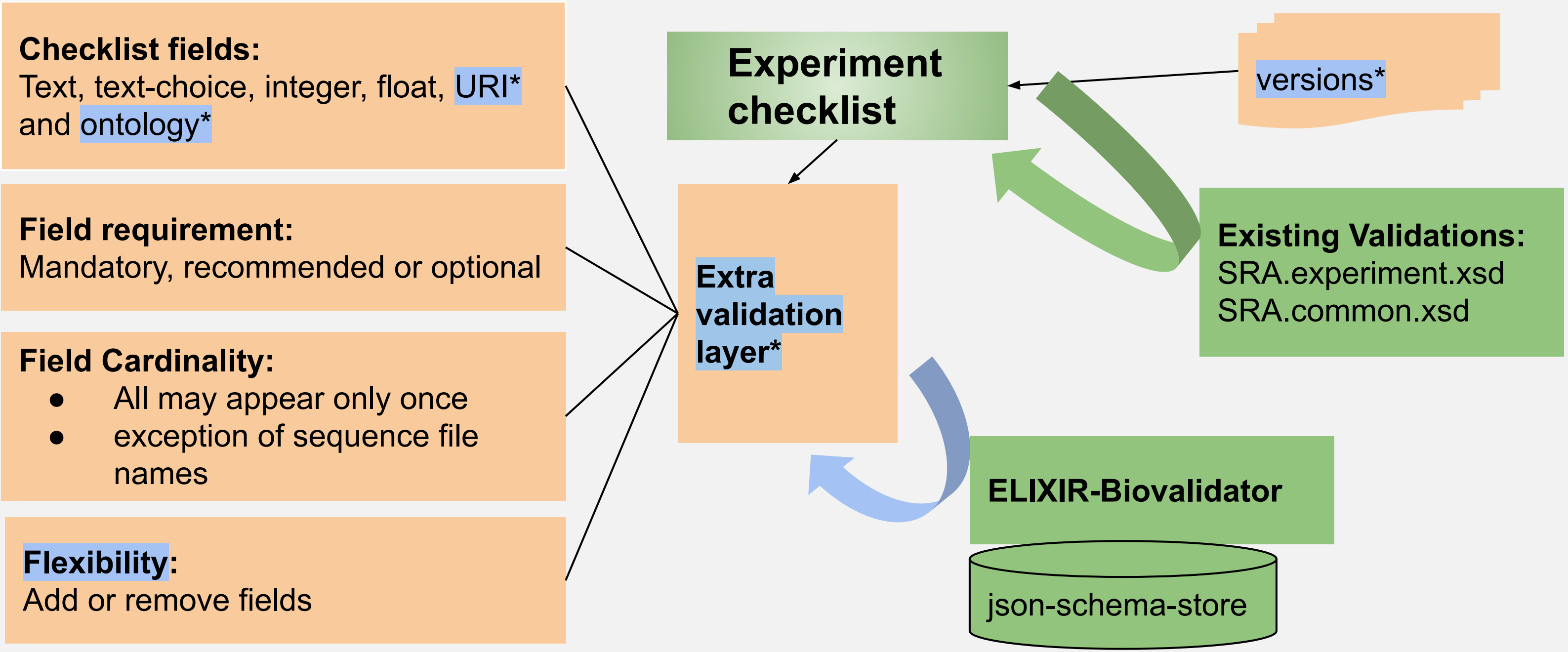
Other Existing Infrastructure

JSON SCHEMA

ELIXIR BIOVALIDATOR (can use from web or local CLI)

Requirements for the new Experimental Checklist

Proposing a checklist system for sequencing experiments which has the same conceptual design as the sample checklist system. However it will have some key improvements*



Checklist fields:
Text, text-choice, integer, float, URI* and ontology*

Field requirement:
Mandatory, recommended or optional

Field Cardinality:

- All may appear only once
- exception of sequence file names

Flexibility:
Add or remove fields

Experiment checklist

versions*

Extra validation layer*

Existing Validations:
SRA.experiment.xsd
SRA.common.xsd

ELIXIR-Biovalidator

json-schema-store

From

To

One size fits all input template

field	value
Organism	soil metagenome
Experiment Accession	ERX2625649
Instrument Platform	LS454
Instrument Model	454 GS FLX Titanium
Center Name	PAU UNIVERSITY
Library Layout	SINGLE
Library Strategy	AMPLICON
Library Source	METAGENOMIC
Library Name	unspecified
Library Selection	PCR

From

To

One size fits all input template

field	value
Organism	soil metagenome
Experiment Accession	ERX2625649
Instrument Platform	LS454
Instrument Model	454 GS FLX Titanium
Center Name	PAU UNIVERSITY
Library Layout	SINGLE
Library Strategy	AMPLICON
Library Source	METAGENOMIC
Library Name	unspecified
Library Selection	PCR

Experiment Type specific template

-with certain fields omitted

-with some fields values pre-filled with the most likely value

additional fields and relevant values in this example:

field	value
Experiment type	METABARCODING
Target loci	16S rRNA

PCR primers	*"pcr_primers": {*
	"fwd_name": "", "fwd_seq": "",
	"rev_name": "", "rev_seq": ""
	}

validation of combinations:

Library Strategy + **Library Source**

Platform + **Model**

European Nucleotide Archive EMBL-EBI

Major Aspects to Implement:

For each experiment type

checklist template

useful experiment fields

checklist schema

Contains the dependencies: data types, between fields

Method of user validating filled out templates

+

Method of submitting templates

Short term: JSON will be converted to XML and validated against SRA_experiment.xml and SRA_common.xml

Documentation of this process and details

Currently the details for each experiment type are automatically generated as md

Example Schema - with built in validator

https://github.com/enasequence/ena-experiment-checklist/blob/main/data/schema/METABARCODING_schema.json

E.g. snippet of a example experiment checklist and the validator

Field	Value
instrument platform	ILLUMINA
instrument model	Illumina HiSeq X"

```
{
      "if": {
        "properties": {
          "instrument_platform": {
            "const": "ILLUMINA"
          }
        }
      },
      "then": {
        "properties": {
          "instrument": {
            "enum": [
              "Illumina HiSeq 4000",
              "Illumina HiSeq 2500",
              "Illumina HiScanSQ",
              "Illumina Genome Analyzer IIx",
              "Illumina MiSeq",
              "Illumina HiSeq X",
              "unspecified",
              "Illumina Genome Analyzer II",
              "Illumina HiSeq 100",
              "Illumina HiSeq 3000",
              "Illumina HiSeq 2000",
```

Anticipating “power” users testing their filled out template against the JSON schema, by using biovalidator

Experiment Types

current list of experiment types: CHROMATIN_RELATED, CHROMOSOME_CONFORMATION_CAPTURE, DNA_BARCODING, EPIGENOMIC, EXOME_SEQUENCING, GENOMIC, GENOTYPING, METABARCODING, METAGENOMIC_SEQUENCING, METATRANSCRIPTOMIC, SPATIAL_TRANSCRIPTOMIC, TRANSCRIPTOMIC, VIRAL_RNA_GENOME

Example

“experiment_type”: “METABARCODING”,
“experiment_type_definition”: “Metabarcoding is the barcoding of DNA/RNA (or eDNA/eRNA) in a manner that allows for the simultaneous identification of many taxa within the same sample. The main difference between barcoding and metabarcoding is that metabarcoding does not focus on one specific organism, but instead aims to determine species composition within a sample.[WIKIPEDIA]“,
“experiment_type_ontology_id”: “EDAM:320",

Trying to use existing ontologies and definitions. Seeking to align with EGA

Current Example (for reference)

Focusing on the Metabarcoding:

- Template: https://github.com/enasequence/ena-experiment-checklist/blob/main/data/output/METABARCODING.json
- Specific doc for each template:

https://github.com/enasequence/ena-experiment-checklist/blob/main/docs/experiment_types/METABARCODING.md

- Schema: https://github.com/enasequence/ena-experiment-checklist/blob/main/data/schema/METABARCODING_schema.json

Input Configuration file

https://github.com/enasequence/ena-experiment-checklist/blob/main/data/input/ExperimentChecklistIn.json

Summary

- Main aim: make the sequence data easier to find and reuse
-
- Near Future:
  - Checklists specific to experiment types
  - Consistency of terms in related fields
  - Users can validate metadata themselves without submitting
  - Using modern technologies: e.g. JSON schemas

Feedback and suggestions welcome

Acknowledgement

Internal

EMBL- EBI ENA Team

- Josie Burgin and Guy Cochrane
- Contents team, especially: Zarah Waheed, Gabi Rinck and Joana Paperio
- Technical team: Rasko Leinonen

Biosample Biovalidator team

- Technical team: Isuru Liynage

EMBL- EBI EGA Team

- Coline Thomas and Marcos Casado

External

GSC and also GA4GH Sequencing Experiment Metadata workstream

- Informative discussions

JSON Schema - it is amazingly elegant and powerful

Project Funders:

1) MGP-IV

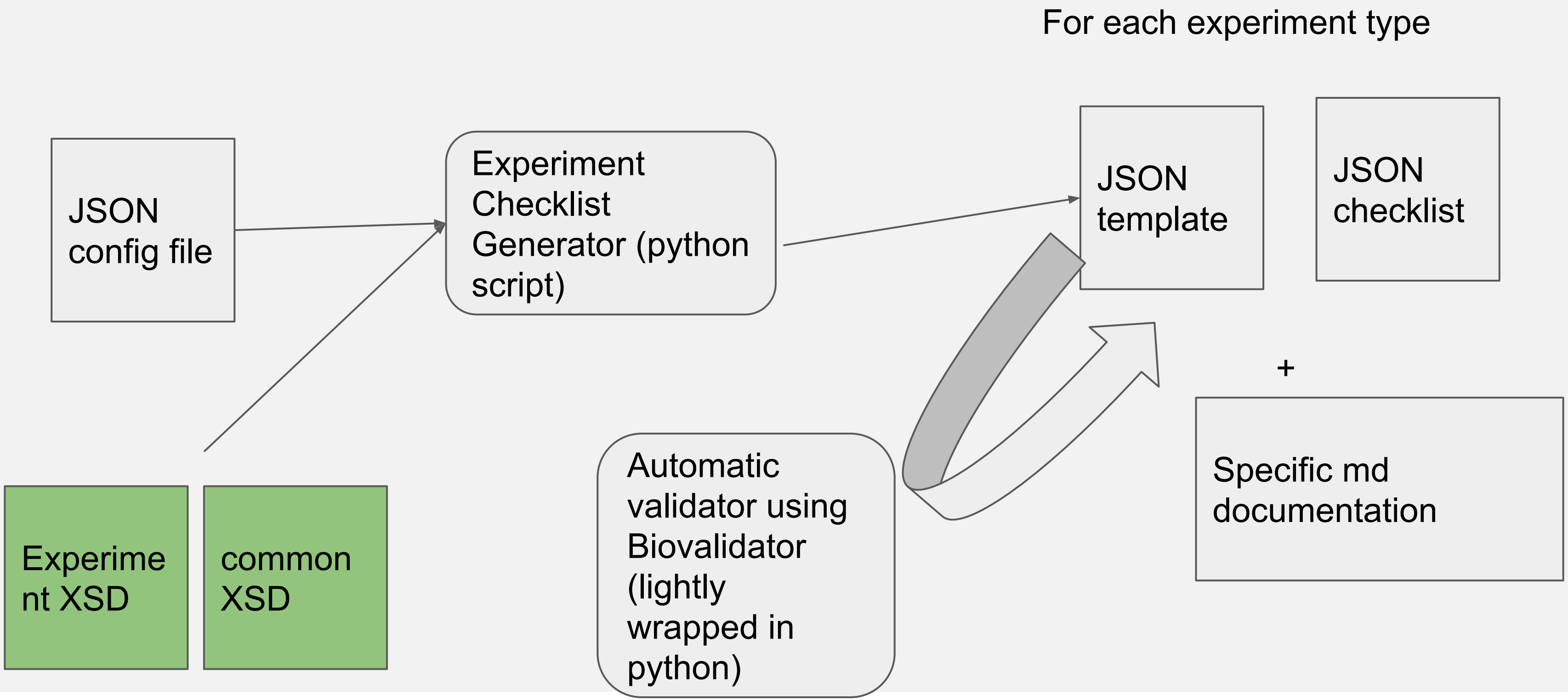
UKRI

Biotechnology and Biological Sciences Research Council

2) **European Advanced Infrastructure for Innovative Genomics**

EASI Genomics

Components in place (alpha release)



For each experiment type

JSON config file

Experiment Checklist Generator (python script)

JSON template

JSON checklist

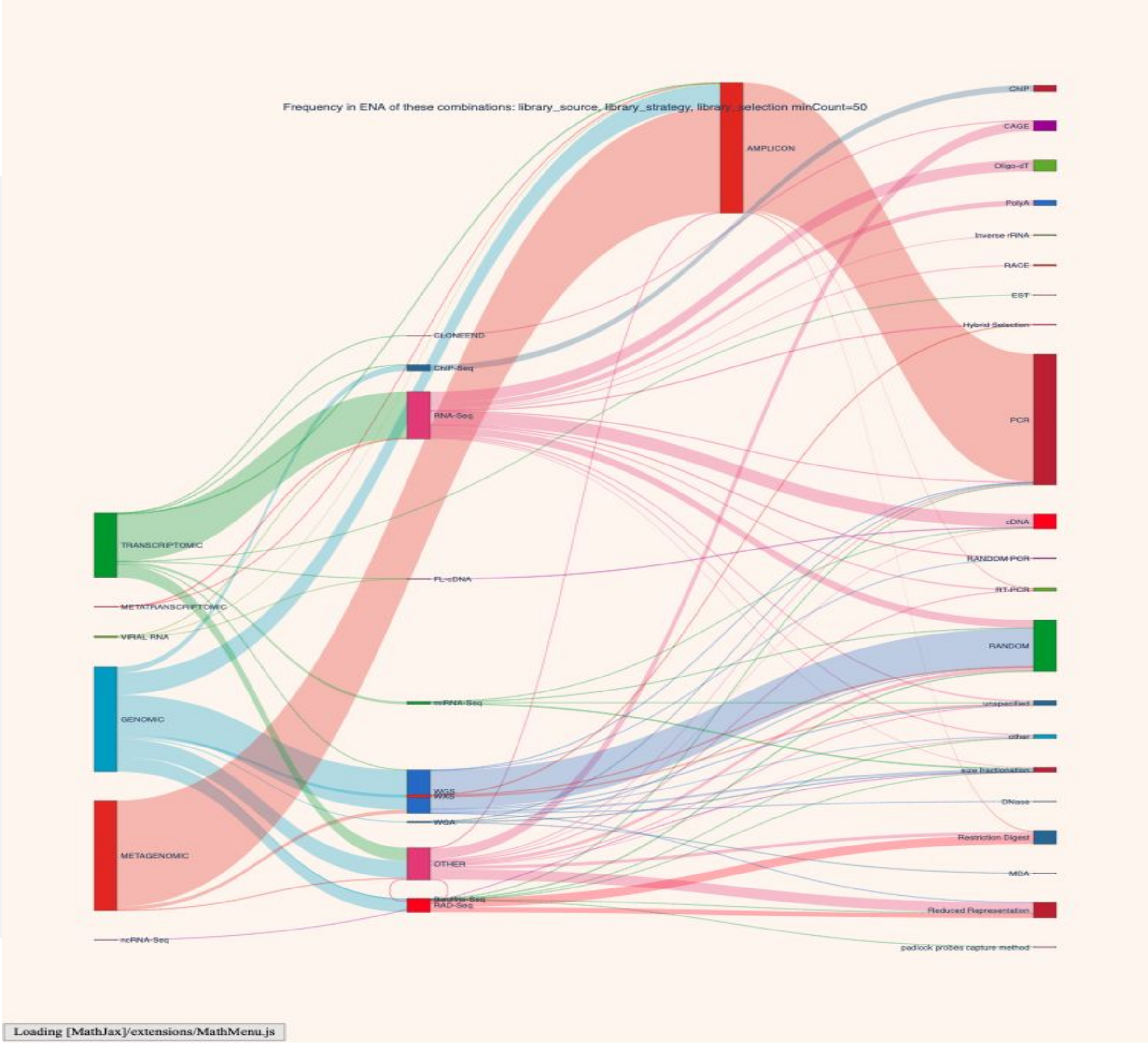
+

Specific md documentation

Experiment XSD

common XSD

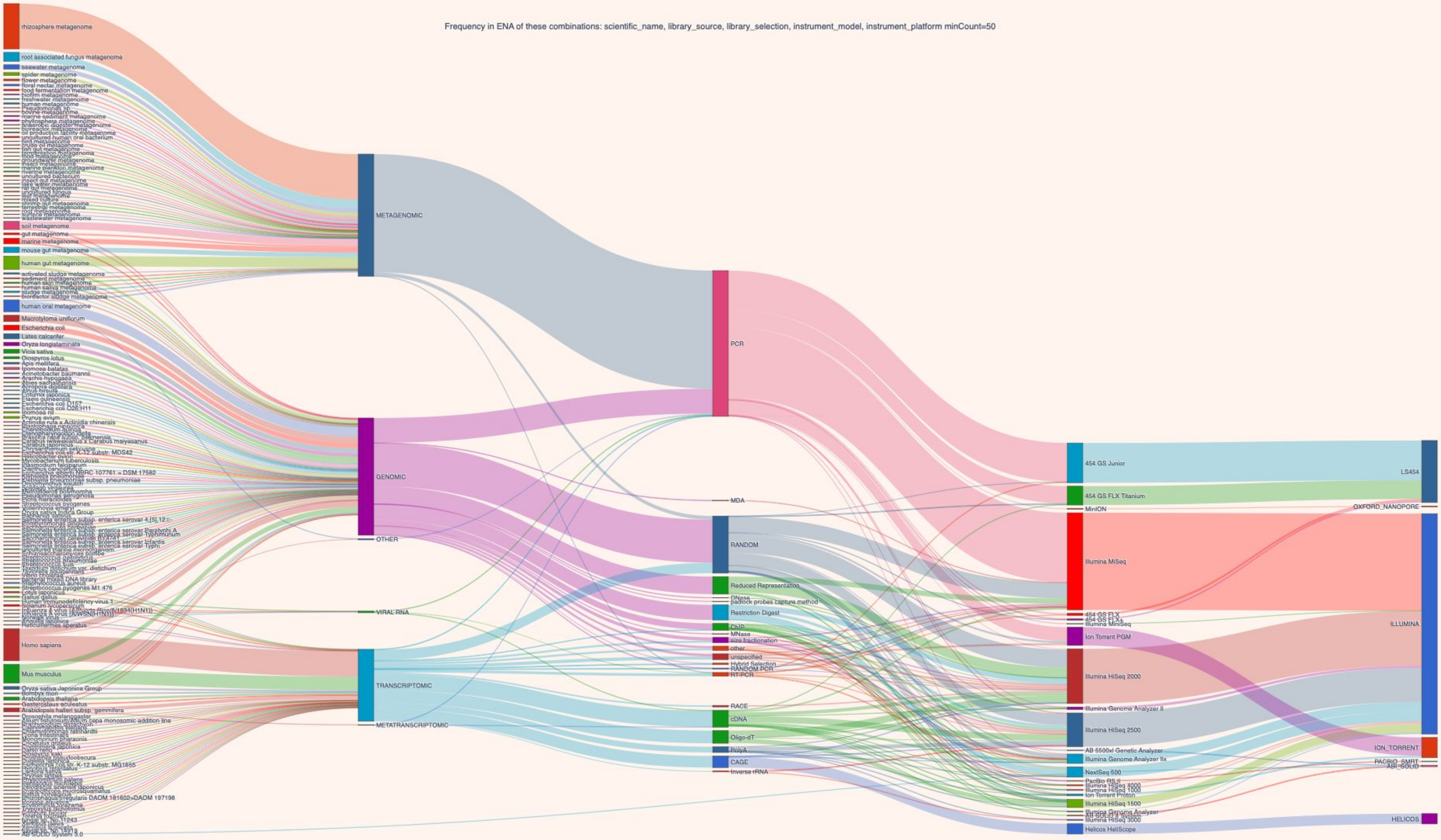
Automatic validator using Biovalidator (lightly wrapped in python)



Frequency in ENA of these combinations: library_source, library_strategy, library_selection minCount=50

Loading [MathJax]/extensions/MathMenu.js

Frequency in ENA of these combinations: scientific_name, library_source, library_selection, instrument_model, instrument_platform minCount=50



Abstract

- Core and diverse sample metadata has been explicitly captured with checklist templates for a number of years, by the European Nucleotide Archive(ENA) and other INSDC partners. There is now a broader and more complex spread of sequencing experiment related metadata that could usefully be collected too, due to the increasing use of sequencing technologies to study the general biological world, particularly for human health and the environment. Capturing experiment metadata information more accurately and consistently will increase the usefulness of the data, by making it more FAIR.
- We are exploring experimental checklists conceptually similar to existing sample level checklists to tailor metadata provided for different 'types' of sequencing experiments. We have integrated learnings from sample checklists, including the need to have checklist versioning and dependency validation. To do the initial validation for the experiment checklists, we are using: JSON, JSON schema and ELIXIR bio validation technologies. These can rapidly catch most validation issues and provide immediate feedback to users. Deeper automated validation will still be performed to ensure INSDC standards.
- Currently, we have a dozen "experiment type" checklists ranging from metabarcoding to spatial transcriptomics. These experiment type checklist JSON and accompanying JSON schema files are all driven from a single JSON configuration file. It will be straightforward and sustainable to add further experiment types.
- A pilot use and submission of experiment type checklists is planned for later this year. All code and documentation is publicly accessible: https://github.com/enasequence/ena-experiment-checklist/
- In this talk, we will outline what we are doing and illustrate how it will improve the standardisation of sequence experimental metadata.