Initiative for a new Research Data Alliance (RDA) working group:

FAIRification of Genomic Tracks WG

– Enabling data-driven life science with sub-dataset granularity through uniform discovery and access of sequence annotation data

Building on: FAIRtracks + omnipy

Sveinung Gundersen, ELIXIR Norway / Centre for Bioinformatics, University of Oslo

A selection of major consortia producing genomic data over the last 20 years

FANTOM

1995 –>

THE CANCER GENOME ATLAS

2005 – 2014

PCAWG
PanCancer Analysis OF WHOLE GENOMES

2014 –>

ROADMAP epigenomics PROJECT

2008 – 2018

BLUEPRINT epigenome

2011 – 2017

ENCODE

2003 –>

International Cancer Genome Consortium

2008 –>

FAANG
Functional Annotation of Animal Genomes

2015 –>

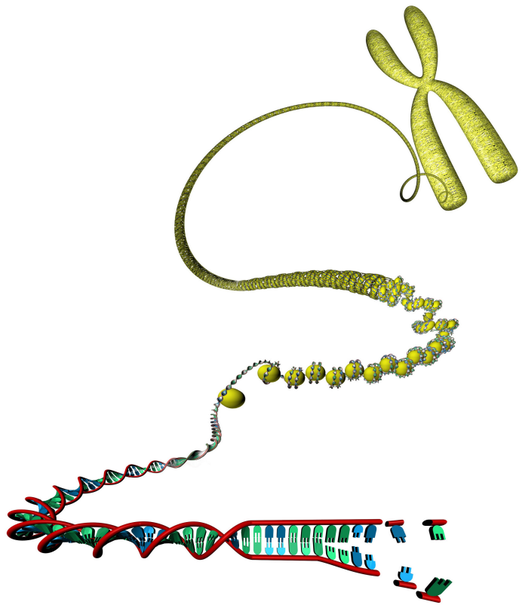
Public data costing billions to produce!

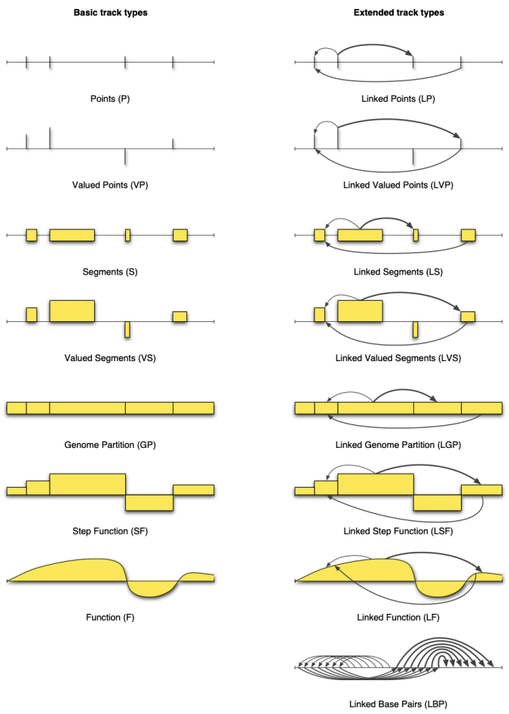
– split into separate data portals with largely incompatible metadata models

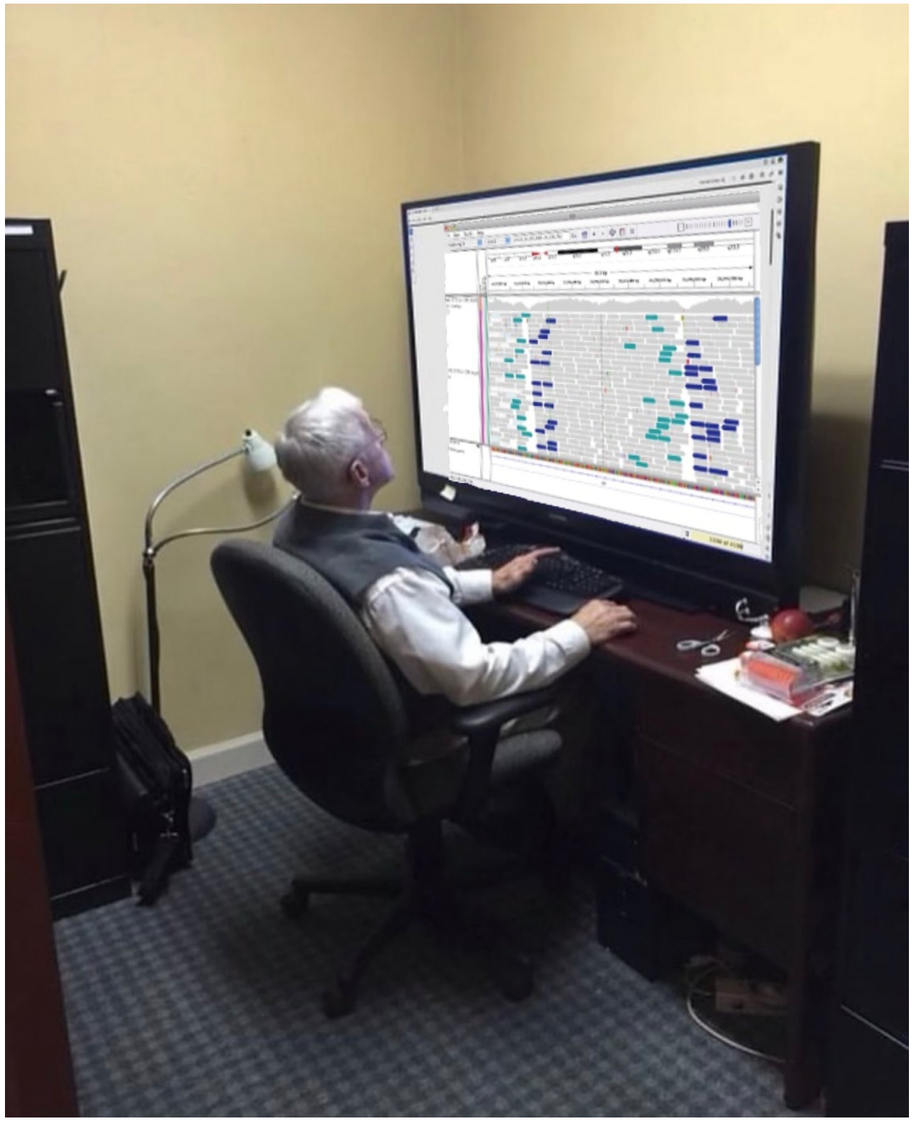
... plus 1000s of smaller scale datasets, many with little metadata

Genomic Tracks

– unified model for data analysis





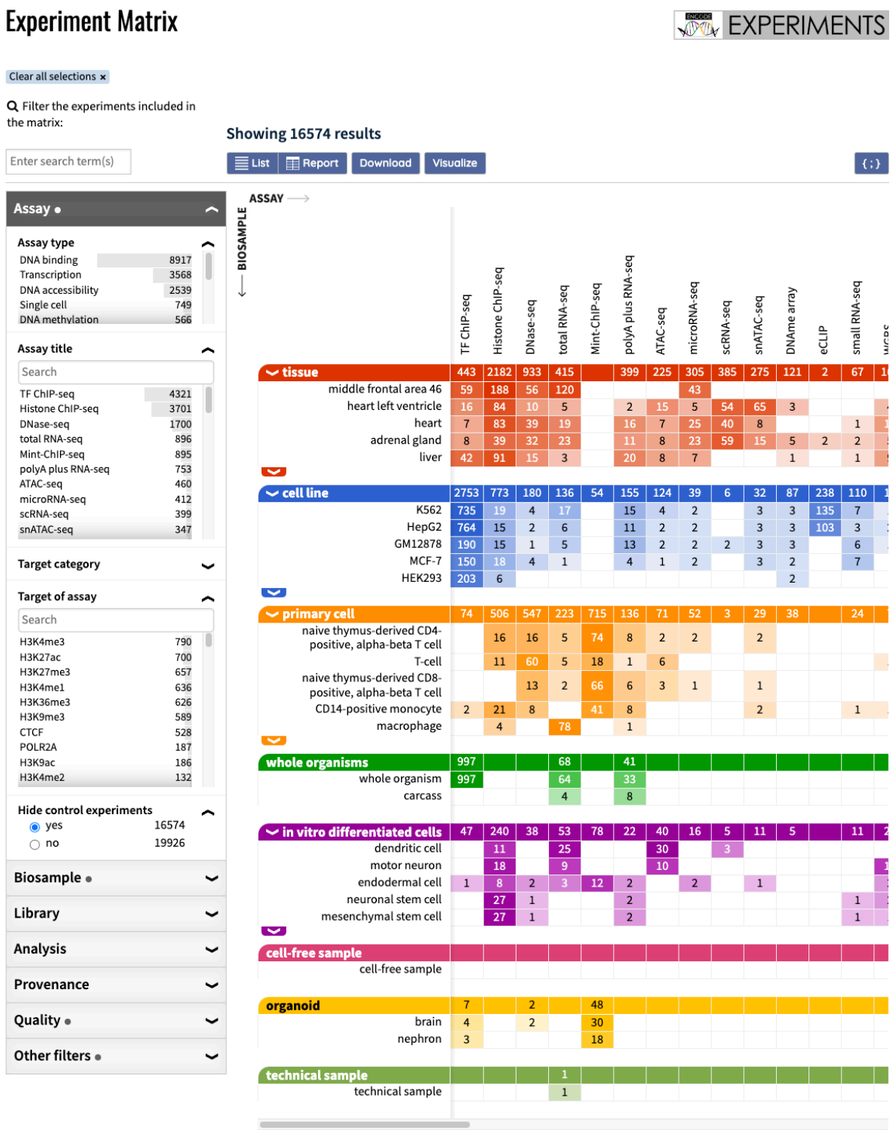


Genome browser meme by Jedidiah Carlson

The mental model of data elements positioned along a reference genome is essential also for non-visual analysis

"Tracks" in this broader sense covers many common file formats, such as BED/BigBED, WIG/BigWIG, SAM/BAM, VCF, GFF, ...

Variation of data content in genomic tracks



- Gene regions, repeating elements, conserved regions
- Chromatin accessibility (e.g., DNase I Hypersensitivity)
- Binding of Transcription Factors to DNA
- Histone modifications along DNA
- Gene expression, Gene fusion, Transcription start sites (TSS)
- Cis-regulatory elements (promoters, enhancers...)
- DNA methylation
- 3D genome structure
- Variation: GWAS SNPs, e-QTLs, SNVs, CNVs, ...
- Virus insertion sites

= Any "omics" data file with reference genome positions!

What is special about track files?

A. Routinely generated through standard pipelines/tools and typically ***stored alongside raw sequence files in larger datasets***

B. Represent ***summaries of the raw data:***

   – A genomic track relates to the raw data much like an abstract describes a scientific publication!

These two properties together makes track files uniquely suited for:

Data-driven discovery of genomic datasets and their relationships – with sub-dataset granularity (at the experiment level)!

FAIR aspects of Genomic Tracks – Potential for Improvements

Findable

- Global identifiers for track files, as well as track collections, studies, samples, and experiments
- Search and import of individual track files across repositories, also repositories not supported by consortia data portals
- Search using formal (non-free-text) queries

Accessible

- Easy (automated) retrieval of track data
- Persistence of track data and versioned metadata

Interoperable

- Lack of standard metadata model with practically useful attributes
- Annotation using community-accepted vocabularies/ontologies
- Cross-references to records in relevant (meta)data repositories

Reusable

- Support for detailed context-specific metadata content together with standardised summary attributes
- Simple process for data providers to submit data and metadata that at the same time accommodates the required stringency for (automated) downstream usability
- Easily available data usage licenses
- Detailed provenance of experimental and *in silico* analysis steps

Gundersen S, Boddu S, Capella-Gutierrez S *et al.* Recommendations for the FAIRification of genomic track metadata [version 1; peer review: 1 approved]. *F1000Research* 2021, **10**(ELIXIR):268 (https://doi.org/10.12688/f1000research.28449.1)

Timeline of the FAIRtracks project

The Genomic HyperBrowser
— statistical analysis of tracks

Manual collection of tracks w/metadata

-> Did not scale

ELIXIR implementation study:

– FAIRification of genomic tracks

ELIXIR NORWAY

ELIXIR SPAIN

EMBL-EBI

Galaxy integration

FAIRtracks.net Web site

omnipy (Meta)data wrangling framework

2010 · · · 2017 — 2018-2020 — 2021 — 2022 — 2023

GSuite HyperBrowser

Prototype of data import tool

– harmonised a few *ad hoc* metadata fields

-> Also did not scale, but sparked the idea of FAIRtracks!

F1000 Research

Publication of recommendations

ELIXIR Recommended Interoperability Resource

RDA RESEARCH DATA ALLIANCE

New RDA working group proposal

Supported by RDA TIGER

"FAIRification of Genomic Tracks WG"
https://tinyurl.com/bddsua24

Main Goals

Allow researchers to:

- Discover and access genomic data
- ... from disparate repositories
- ... from different types of experiments
- ... at a sub-dataset granularity
- ... through categorical search in uniformly FAIRified metadata that relate to "genomic tracks"

Ulterior goal

Enable data-driven discoveries through analysis of condensed and heterogeneous genomic data files from disparate repositories

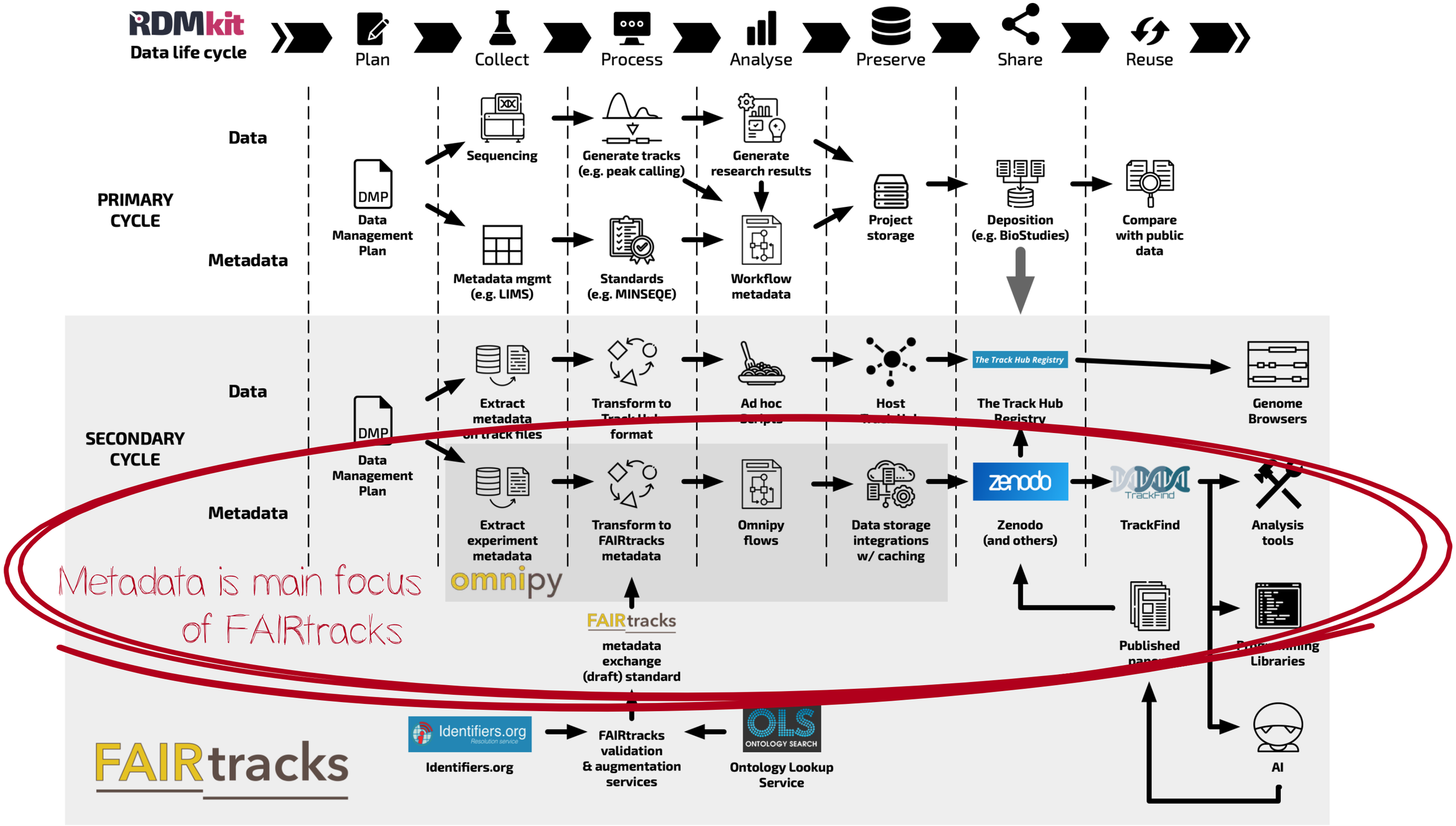
Detailed Goals

- **Build a global community, bringing in key people and entities to build a solid foundation for a truly global standard**
- **Further development of recommendations / metadata schemas**
  - Refine ontology selections, metadata fields
  - Support more experiment types, species, biospecimen types etc. based on expert knowledge
- **Integrate with new data repositories or other data sources**
  - Implement continuous metadata transformation pipelines from different collections containing track files
- **Integrate with downstream software tools and frameworks**
- **Support use cases from analytical end users, e.g. AI-driven methodologies**
- **Build on other efforts:**
  - Adopt recommendations/standards from RDA, GA4GH etc.
  - Integrate with relevant solutions: metadata standards, metadata and ontology term mapping, etc.

What we have (1/4)

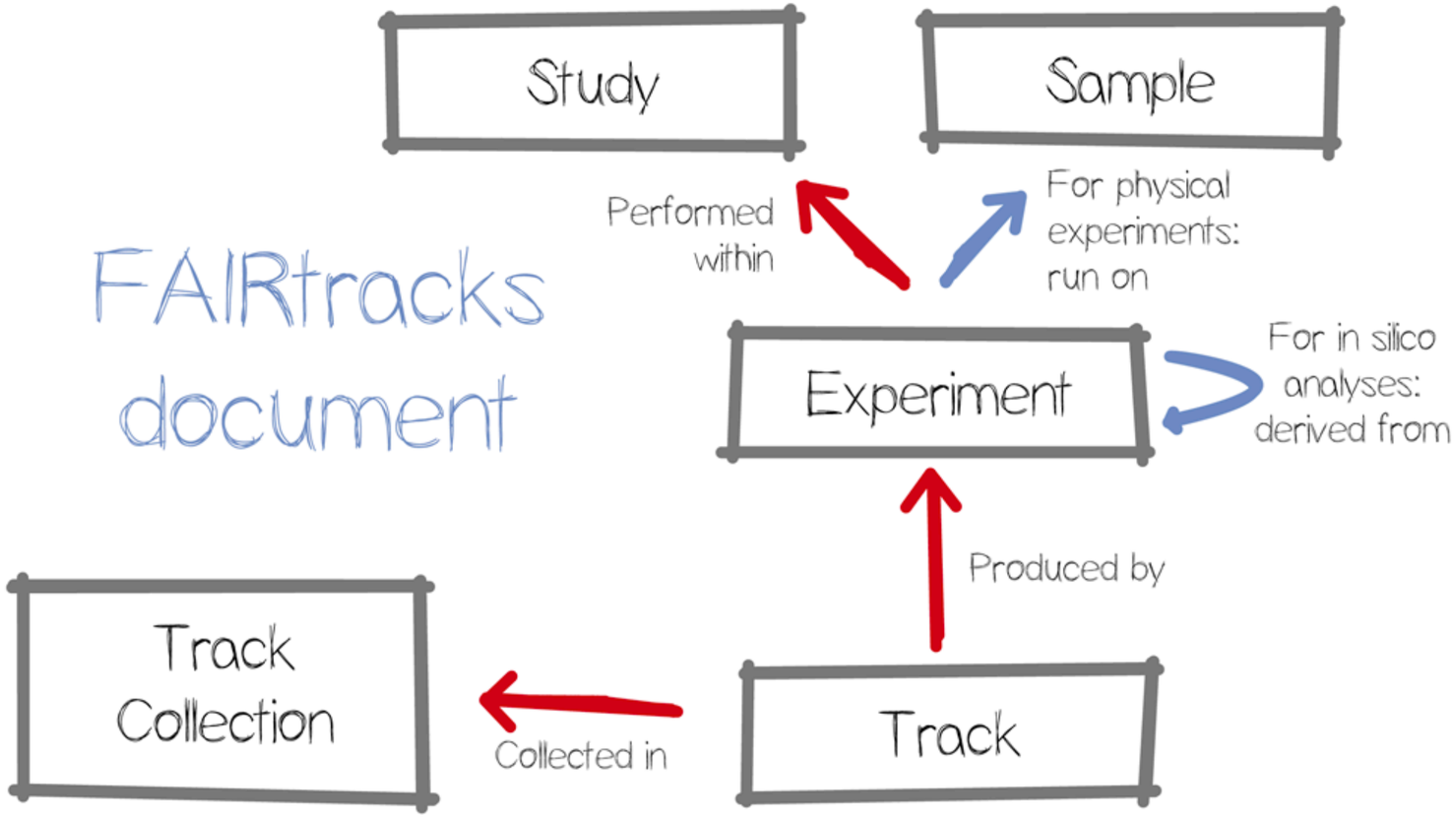
FAIRtracks

- **FAIRtracks draft metadata exchange standard**
- **Services for augmentation and validation of FAIRtracks metadata**
- **TrackFind**
  - categorical search of track metadata (web + REST API)
- **User–friendly data import tool in Galaxy**
  - using TrackFind as backend



Icons: RDMkit, Google Material Symbols, Flaticon.com

FAIRtracks draft metadata standard



- **Minimal metadata exchange format**
  - Designed for minimal information needed for downstream usability
- **JSON Schema-based validation**
- **Designed for interoperability:**
  - identifiers.org/N2T.net – resolvable CURIE references to external records
  - use of ontology terms as much as possible
  - currently JSON, but will probably be moved to JDON-LD for semantic web, RO-Crate, etc.
- **5 main object types, mappable to other common standards**
- https://fairtracks.net/standards

What we have (2/4)

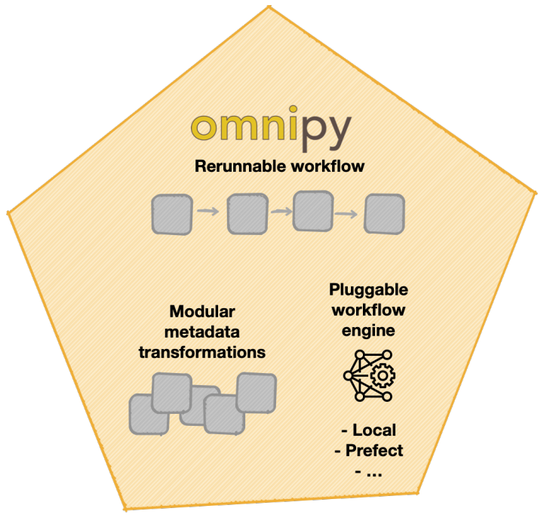
omnipy

- **High-level Python library for type-driven data wrangling and scalable data flow orchestration**

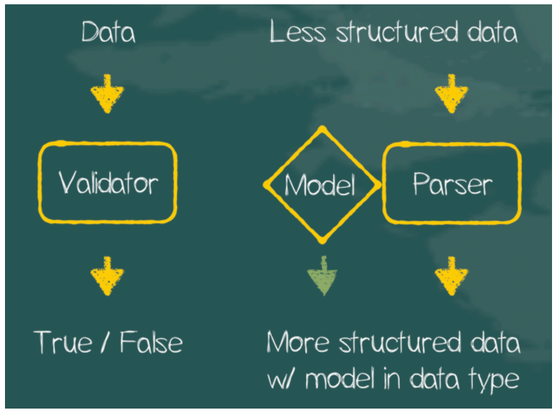
- **Developed to allow for rerunnable metadata transformation flows that:**
  - Help researchers extract, manipulate and integrate data and/or metadata from different sources
  - Supports conversion of existing track metadata to the FAIRtracks standard
  - Is general enough to support other data transformation flows, even in other domains

omnipy

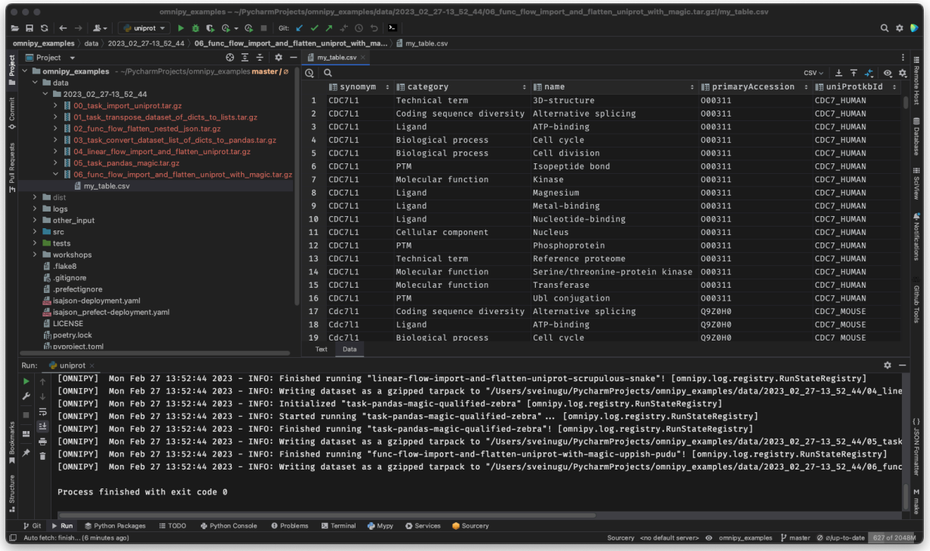
- **Import (meta)data in almost any shape or form**
  - nested JSON
  - tabular (relational) data
  - binary streams
  - other data structures
- **Set up a step-by-step data flow that:**
  - Reformats data structures
  - Cleans up errors
  - Removes duplicate data
  - Maps and integrates dataset fields
- **Provide a catalog of generic task and flow templates that the researcher can refine according to the use case**
- **For large datasets:**
  - Omnipy allows local test jobs to be seamlessly scaled up to the full dataset and offloaded to external compute resources



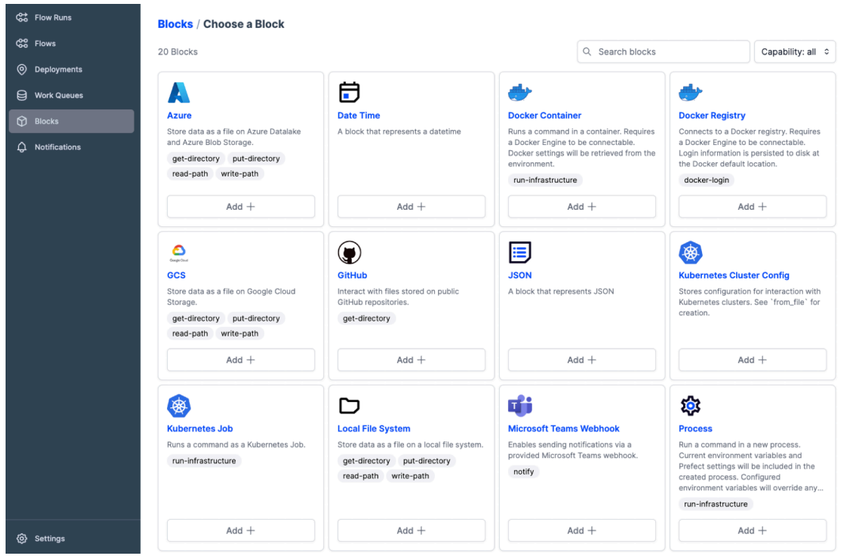
Highly modular software architecture



"Parse, don't validate"

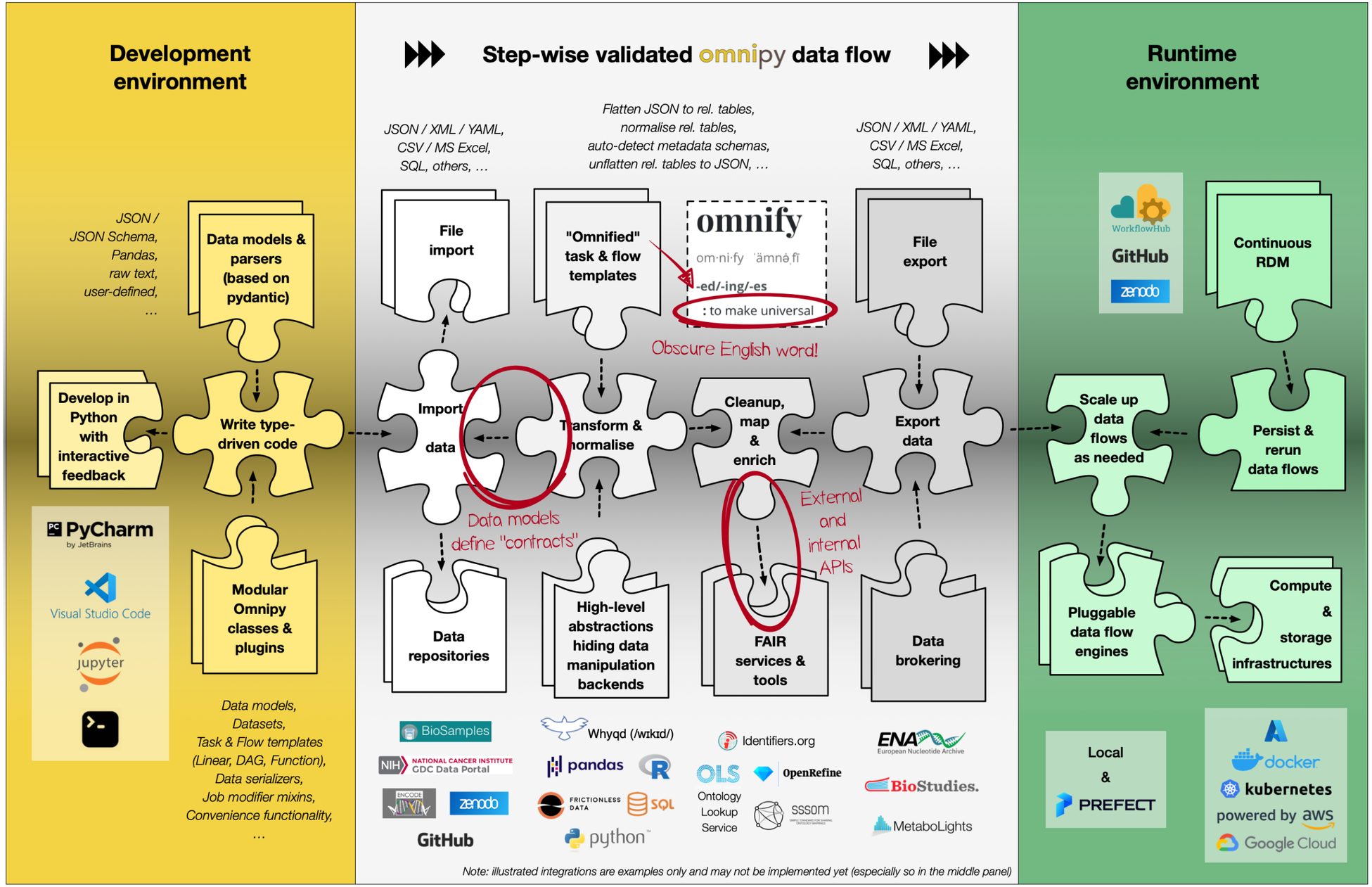


Develop, inspect and deploy directly from IDE



Prefect orchestration GUI for local/remote deployment

omnipy — The interoperability layer of data wrangling!



Modular Python library for developing and orchestrating scalable (meta)data flows

What we have (3/4)

FAIRtracks.net website

- Information on:
  - Genome browsers and genomic tracks
  - FAIR interoperability solutions
  - FAIRtracks-related services
  - Background material
- Overview tables:
  - Genome browsers - services and software
  - Data portals and other repositories hosting track data
  - Ontologies, datasets, and controlled vocuabularies
- *... plus more to come*
- Aims to be a hub for community involvement

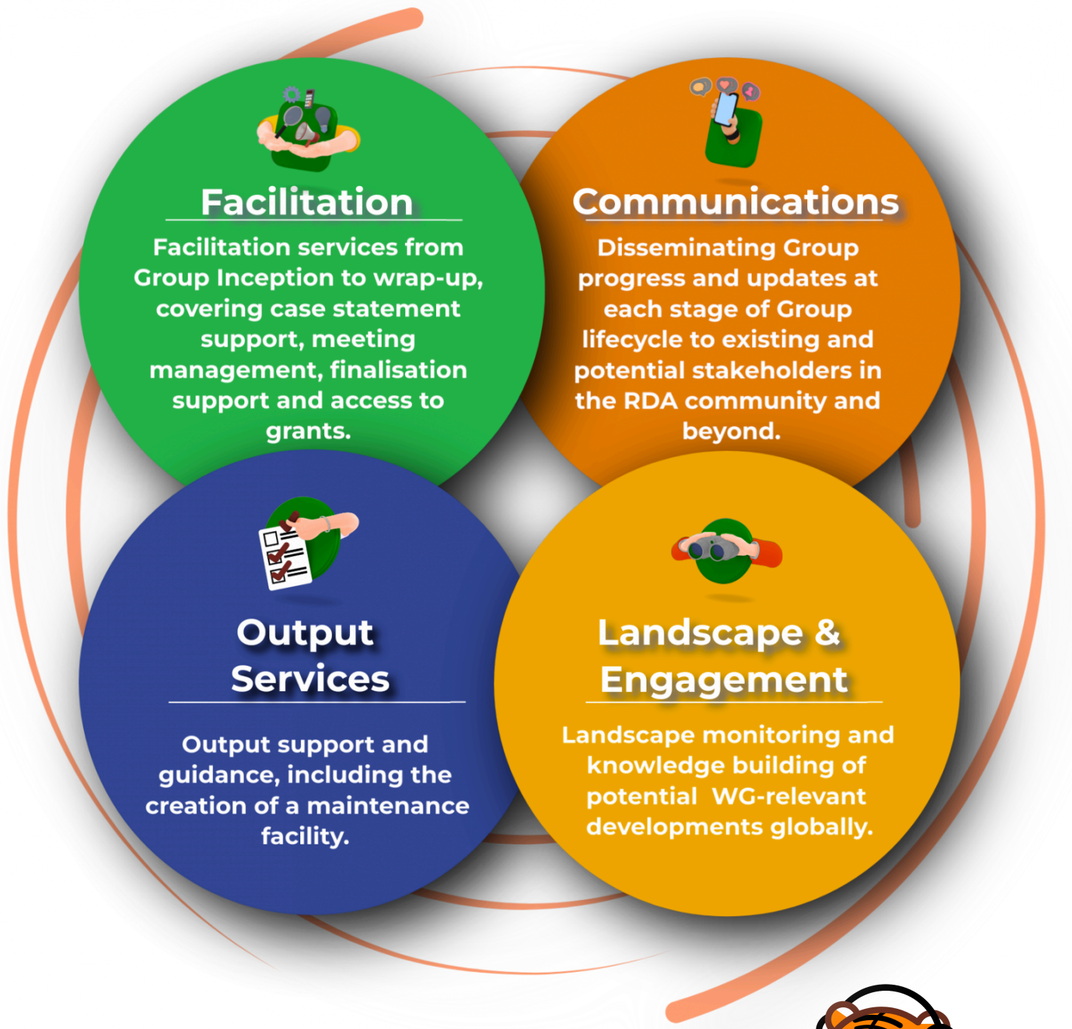
- **Collaboration between**
  - ELIXIR Norway
  - ELIXIR Spain
  - EMBL/EBI

- **Existing and planned integration efforts with:**
  - Genomic HyperBrowser
  - EPICO (BLUEPRINT data analysis portal)
  - Track Hub Registry
  - ENSEMBL Genome Browser
  - Euro-FAANG project as use case through Peter Harrison (EMBL/EBI)

FAIRification of Genomic Tracks WG

What we have (4/4)

- **Support from RDA TIGER project**
  - "Research Data Alliance facilitation of Targeted International working Groups for EOSC-related Research solutions"
  - Kick-off Jan 2023
  - Funded through HORIZON-INFRA-2022-EOSC-01-04 call
- **FAIRification of Genomic Tracks WG has been selected as a pilot to receive support/facilitation**
- **We still need to write a case statement to be evaluated by the RDA Technical Advisory Board!**



What we wish for

- **Global members of a FAIRtracks WG, preferably:**
  - Data producers
  - FAIR / RDA community
  - Tool developers
  - Analytical end-users / Domain experts
- **WG Chairs outside of Europe!**
- **Help developing the WG case statement**
- **Lots of members and lots of input!**
- **FAIRtracks WG kick-off in Salzburg, Oct 2023?**

Interested in contributing to the working group?

- **Register you interest by**
  - Filling out our survey:
    - FAIRification of Genomic Tracks: Community interest for an RDA working group (https://nettskjema.no/a/fairtracks-wg)
  - and/or send an email to:
    - fairtracks@elixir.no

- **Working document for the "FAIRification of Genomic Tracks WG" initiative**
  - https://tinyurl.com/bddsua24

- **Shareable link to this presentation**
  - https://tinyurl.com/2xhhwsb8

Acknowledgements

FAIRtracks

The FAIRtracks team

+

EBI/EMBL, Hinxton, UK

ELIXIR Norway,
Centre of Bioinformatics, UiO

ELIXIR Spain,
Barcelona Supercomputing Centre

ELIXIR Norway,
NTNU & UiB

+

+

Jeanne Cheneby,
Pável Vázquez

Acknowledgements

inb Spanish National Bioinformatics Institute

With funding from The Research Council of Norway

NTNU Norwegian University of Science and Technology

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

EMBL-EBI

UiO : University of Oslo

elixir NORWAY

RDA TIGER

eosc

elixir SPAIN

Extra slides

Relevant metadata standards

- We have considered important existing standards, authorities, and consumers of track (and related) metadata:
  - INSDC
  - IHEC
  - FAANG
  - DATS (BioCADDIE, GDC)
  - ICGC/PCAWG
  - ISA Model
  - TrackHub file format
  - Zenbu
  - GSuite HyperBrowser
  - ...
- In contrast to many other standards, FAIRtracks have been driven mainly by end-user functionality
- Our focus has *not* been to make sure all important metadata is archived properly
- Rather, FAIRtracks is a **metadata exchange** standard!

Identifiers for genomic tracks?

- Should there be globally unique, persistent identifiers for genomic track files (e.g. BigBED/BigWIG, VCF, etc)

- No such thing exists, but we highly recommend that they should be created and indexed

- Also, identifiers for *track collections* would be very useful
  - Would easily FAIRify "Mix-and-match" track file collections analysed in particular research papers, if the track files are already FAIRified

Use of ontology terms

- **Advantages:**
  - Interoperability
  - Make use of existing expert knowledge
- **Disadvantages:**
  - Selecting proper ontologies
  - Missing terms – need to coordinate with ontologies
    - But can create simple vocabulary in the meantime
  - Ontology versioning
    - How to update metadata when ontologies change?
    - Our solution: data flows that can be automatically in a CI/CD framework

Metadata augmentation

- Makes machine-readable metadata understandable by humans
  - Fills out ontology labels from term IDs
  - Fetches info by ID in other databases
  - Creates summary fields
  - Other housekeeping tasks

README.md

The FAIRtracks augmentation service

The FAIRtracks draft standard is a set of JSON Schemas that define a minimal standard for genomic track metadata. The FAIRtracks augmentation service is a simple Flask-based service written in Python that provides a HTTP-based API for augmenting a minimal FAIRtracks JSON document with automatically generated content for all the properties defined in the FAIRtracks JSON schemas with `augmented=true.`

The FAIRtracks augmentation service fills out the following fields:

- The newest versions of all ontologies required by the FAIRtracks standard are fetched and the versioned URLs are added to the `doc_info` -> `doc_ontology_versions` object.
- All `term_value` properties are generated by the related `term_id` property by lookup in the relevant ontology.
- `track` -> `file_name` is generated from the `track` -> `file_url` property.a
- The two properties `sample` -> `sample_type` ->summary and `experiment` -> `target` -> `summary` are generated based upon the relevant rule as defined by `sample` -> `biospecimen_class` -> `term_id` and `target` -> `technique` -> `term_id` , respectively, as described in the top-level FAIRtracks schema.
- `sample` -> `species_name` is generated by a lookup in the NCBI Taxonomy Database

Experiment object example – augmented version


```
{
    "global_id": "ega.experiment:EGAX00001215632",
    "local_id": "ERX547964",
    "study_ref": "EGAS00001000326",
    "sample_ref": "S00B1LH1",
    "technique": {
        "term_id": "http://purl.obolibrary.org/obo/OBI_0002017",
        "term_label": "histone modification identification by ChIP-Seq assay"
    },
    "target": {
        "sequence_feature": {
            "term_id": "http://purl.obolibrary.org/obo/SO_0001410",
            "term_label": "experimental_feature"
        },
        "summary": "experimental_feature (Input)",
        "details": "Input"
    },
    "lab_protocol_description":
      "http://www.blueprint-epigenome.eu/index.cfm?p=7BF8A4B6-F4FE-861A-2AD57A08D63D0B58",
    "compute_protocol_description":
      "http://dcc.blueprint-epigenome.eu/#/md/chip_seq_grch38"
},
```


Automatically augmented for human readability

Experiment object example – minimal version

```
{
    "global_id": "ega.experiment:EGAX00001215632",
    "local_id": "ERX547964",
    "study_ref": "EGAS00001000326",
    "sample_ref": "S00B1LH1",
    "technique": {
        "term_id": "http://purl.obolibrary.org/obo/OBI_0002017"
    },
    "target": {
        "sequence_feature": {
            "term_id": "http://purl.obolibrary.org/obo/SO_0001410"
        },
        "details": "Input"
    },
    "lab_protocol_description":
      "http://www.blueprint-epigenome.eu/index.cfm?p=7BF8A4B6-F4FE-861A-2AD57A08D63D0B58",
    "compute_protocol_description":
      "http://dcc.blueprint-epigenome.eu/#/md/chip_seq_grch38"
},
```

Minimal version only requires ontology term identifiers

Overview of metadata fields for Example schema

Experiment

Property	Type	Required
@schema	`string`	Optional
global_id	`string`	Optional
local_id	`string`	Required
study_ref	`string`	Required
sample_ref	`string`	Optional
aggregated_from	`array`	Optional
technique	`object`	Required
target	`object`	Required
lab_protocol_description	`string`	Optional
compute_protocol_description	`string`	Optional
Additional Properties	Any	Optional

AnyOf

Technique

Property	Type	Required
term_id	`string`	Required
term_label	`string`	Optional

Subclass of:
- "Planned process" (OBI ontology) or
- "Operation" (EDAM ontology)

Target

Property	Type	Required
sequence_feature	`object`	Optional
gene_id	`string`	Optional
gene_product_type	`object`	Optional
macromolecular_structure	`object`	Optional
phenotype	`object`	Optional
details	`string`	Optional
summary	`string`	Optional

Example schema: technique -> target dependencies

technique	target sequence_feature	gene_product_type	gene_id	macromolecular_structure	phenotype
bisulfite sequencing assay	open_chromatin_region				
DNase I hypersensitive sites sequencing assay	DNaseI_hypersensitive_site				
histone modification identification by ChIP-Seq assay	(any)				
microRNA profiling by high throughput sequencing assay		MicroRNA			
transcription factor binding site identification by ChIP-Seq assay		Transcription Factor	(any)		
RNA-seq assay		Messenger RNA			
Hi-C assay				Chromatin Structure	
GWAS study					(any)

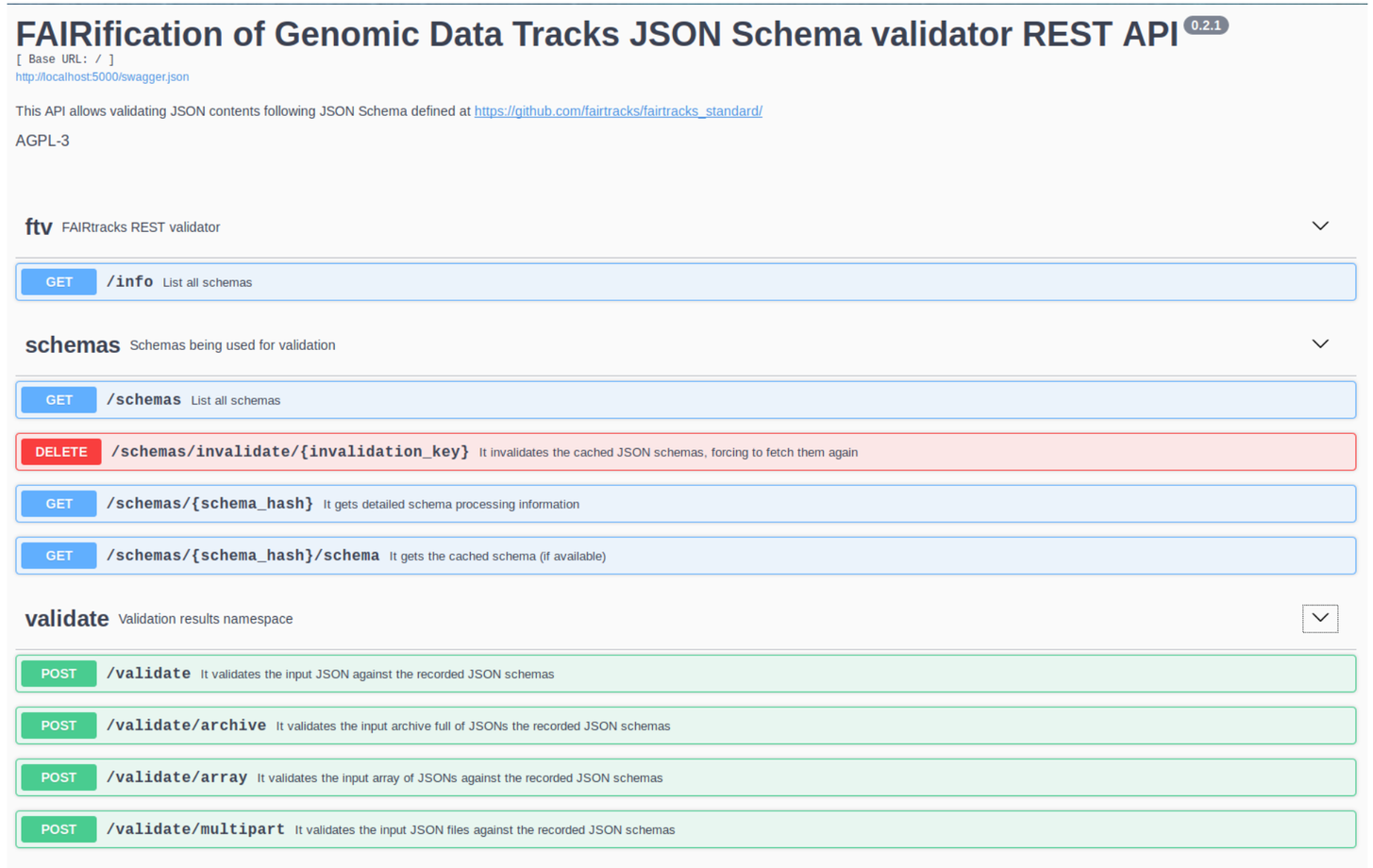
In addition:

- details – text field
- summary – autogenerated based on other fields (incl details)

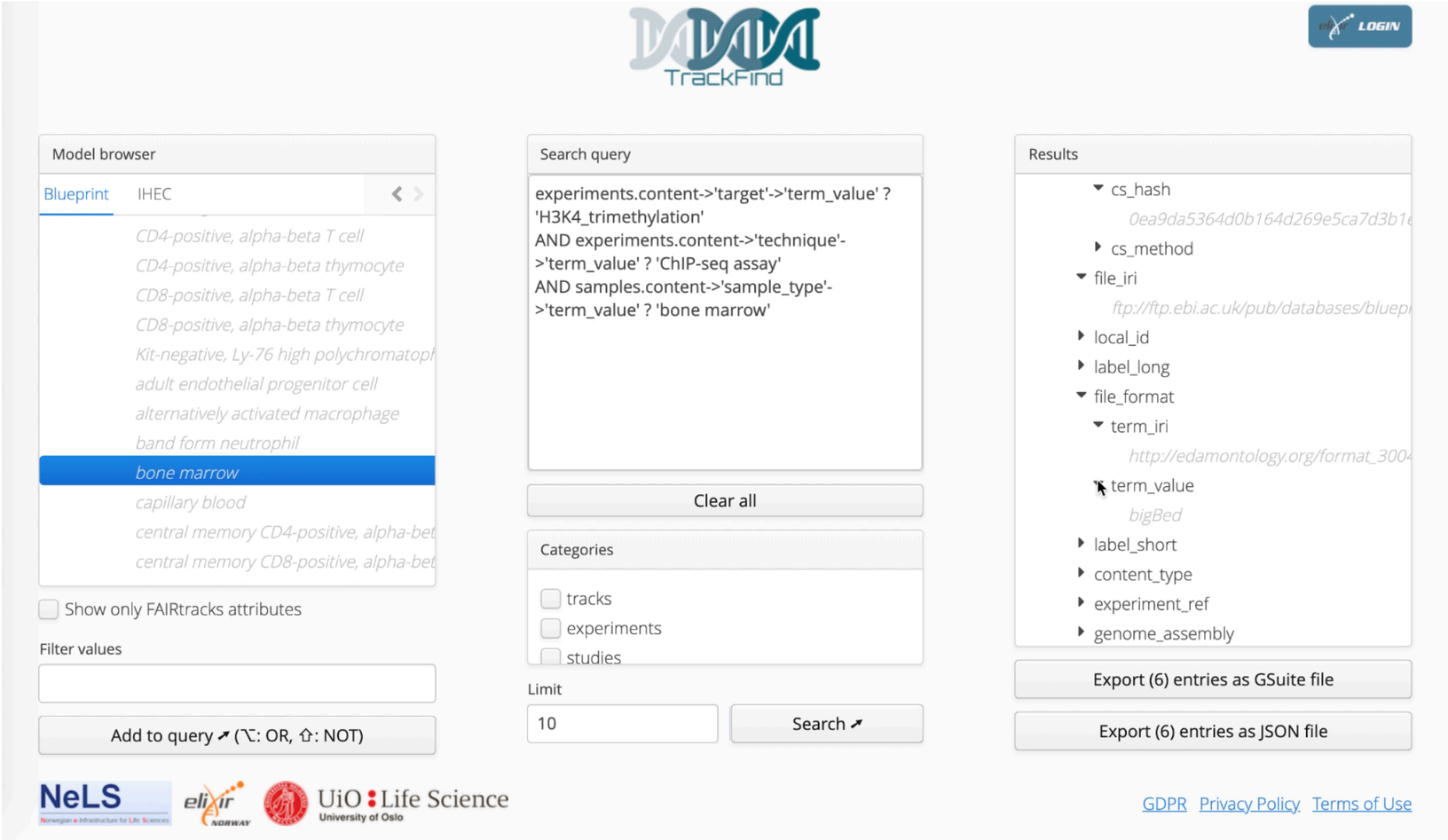
Metadata validation

- Extension of JSON Schema validator developed through OpenEBench

- Validates:
  - identifier.org CURIEs
  - ontology terms
  - duplicate records in relational schemas



TrackFind


TrackFind
LOGIN
Model browser
Blueprint
IHEC
CD4-positive, alpha-beta T cell
CD4-positive, alpha-beta thymocyte
CD8-positive, alpha-beta T cell
CD8-positive, alpha-beta thymocyte
Kit-negative, Ly-76 high polychromatoph
adult endothelial progenitor cell
alternatively activated macrophage
band form neutrophil
bone marrow
capillary blood
central memory CD4-positive, alpha-bet
central memory CD8-positive, alpha-bet
Show only FAIRtracks attributes
Filter values
Add to query (⌥: OR, ⇧: NOT)
Search query
experiments.content->'target'->'term_value' ? 'H3K4_trimethylation'
AND experiments.content->'technique'->'term_value' ? 'ChIP-seq assay'
AND samples.content->'sample_type'->'term_value' ? 'bone marrow'
Clear all
Categories
tracks
experiments
studies
Limit
10
Search
Results
cs_hash
0ea9da5364d0b164d269e5ca7d3b1
cs_method
file_iri
ftp://ftp.ebi.ac.uk/pub/databases/bluep
local_id
label_long
file_format
term_iri
http://edamontology.org/format_3004
term_value
bigBed
label_short
content_type
experiment_ref
genome_assembly
Export (6) entries as GSuite file
Export (6) entries as JSON file
NeLS
elixir NORWAY
UiO : Life Science
University of Oslo
GDPR
Privacy Policy
Terms of Use

Tool integration

- TrackFind client implemented in GSuite HyperBrowser:
  - https://hyperbrowser.uio.no/trackfind_test (search for tool "trackfind")
- JSON and GSuite (http://gtrack.no) formats as metadata / search result exchange format
- Search results can be transferred to the HyperBrowser server, preprocessed, and used in statistical analyses
- EPICO integration still in development