

BioDiv-FAIR-Checker: FAIR metrics implementation for biodiversity resources

Alban Gaignard – CNRS

 Bioschemas community meeting

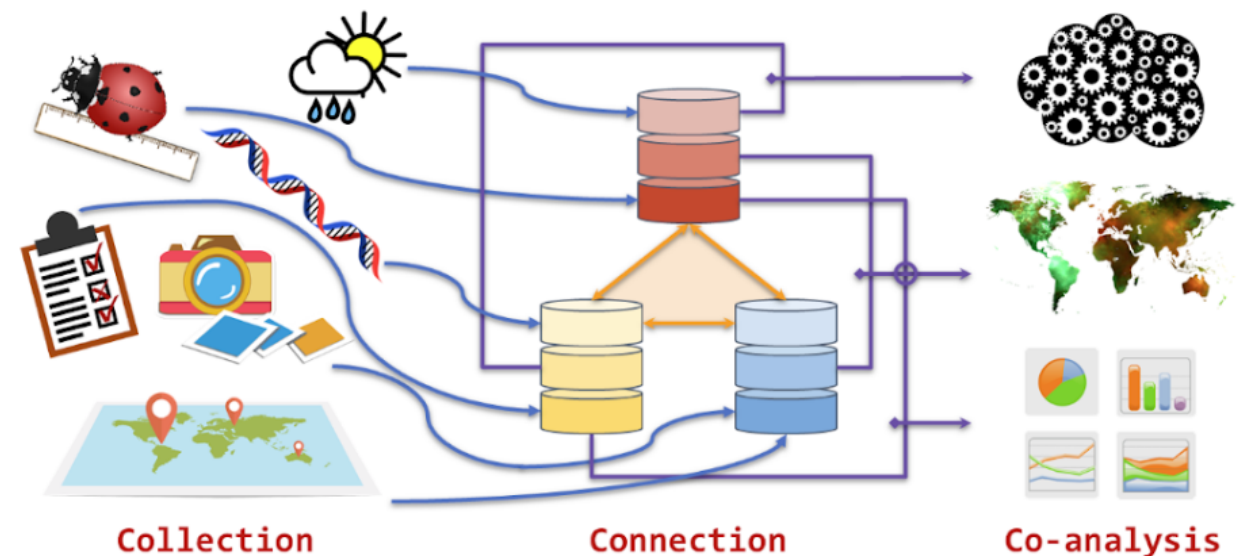
21th of September 2026



community

Elixir BFSP community

- ▶ **Biodiversity**, Food Security and Pathogens community
- ▶ mobilize genetics, phonemics, observation, and omics data
- ▶ focus on research data management (brokering, interop) and analysis

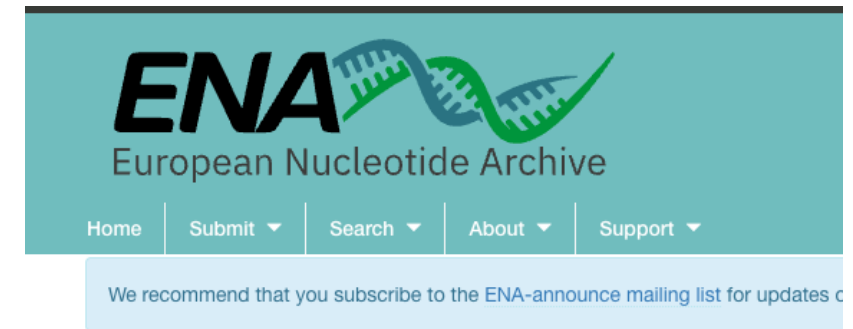


ELIXIR BFSP Strategy: <https://zenodo.org/records/15133254>

Many mature data repositories

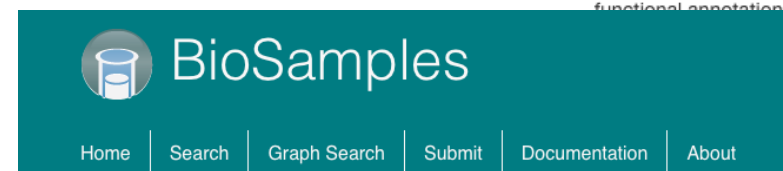
- ▶ Well curated resources
- ▶ We expect good FAIR assessment scores

Raw nucleotide sequences



European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the functional annotation. [More about ENA.](#)

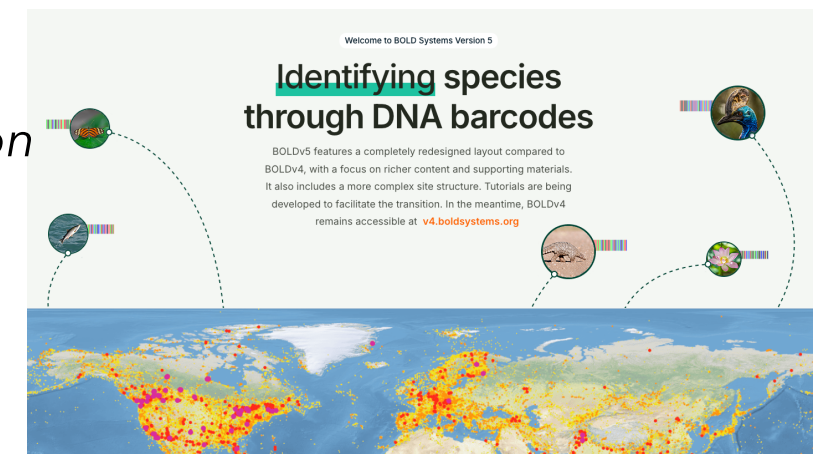


BioSamples an ELIXIR Core Deposition Database, stores and supplies descriptions and industry. Samples are either 'reference' samples (e.g. from 1000 Genomes, HipSci) publicly available data in the European Nucleotide Archive (ENA) or ArrayExpress. BioSamples supports links between sample records and any sample-derived datasets, Archive (ENA) or ArrayExpress, -omics-based datasets such as those in PRIDE or Met samples and accepts direct submissions of sample information.

Sample metadata

BOLD SYSTEMS

Species identification meta-barcoding



Research infrastructure, catalogue of datasets

...

Well curated resources



Object ID: caos-cloud.linode-us-east.25_219.71749d66-f77c-4d7c-a963-a231c06ad99e.jpg

[Image](#)

[Metadata](#)

File Name: CHOLH/RMNH_INS_549233_dor+1328644036.jpg

Orientation: Dorsal

License Holder:

License Year: 2012

License: CreativeCommons - Attribution Non-Commercial Share-Alike

Institution: Naturalis, Biodiversity Centre

Photographer: Naturalis DNA barcoding facility

Identifiers

Process ID: CHOLH011-12 **Sample ID:** RMNH.INS.549233

Record Created: 2012-02-24 **Museum ID:** RMNH.INS.549233

Collection Code: **Field ID:**

Deposited In: Naturalis Biodiversity Center

Associated Datasets:

Specimen Linkout:

Checksum: 9ff7c9dff0b05594fa152d4c0d6c1c8e



Taxonomy

Kingdom: Animalia **Subfamily:** Cholevinae

Phylum: Arthropoda **Tribe:** Cholevini

Class: Insecta **Genus:** *Catops*

Order: Coleoptera **Species:** *Catops picipes*

Family: Leiodidae **Scientific Name Authorship:**

BIN ID: [BOLD:AAF2084](#) [Subspecies:](#)

Identification Method:

Taxonomy Notes:

* Barcode Index Numbers (BIN): Clustered barcode sequences that create OTUs (operational taxonomic units) closely reflective of species groupings

However ...

- FAIR scores for multiple biodiversity resources with FAIR-Checker:

Source	Type of entries	Metadata access point	URL	FAIR score
EBI Biosamples	BioSamples structured	JSON-LD export	https://www.ebi.ac.uk/biosamples/samples/SAMEA112797446.ldjson	45.8 %
EBI Biosamples	BioSamples entry	HTML landing page	https://www.ebi.ac.uk/biosamples/samples/SAMEA112797446	16.7 %
ENA	ENA browser entry	HTML landing page	https://www.ebi.ac.uk/ena/browser/view/SAMEA112797446	16.7 %
ERGA	ERGA portal organism	HTML landing page	https://portal.erga-biodiversity.eu/organism/SAMEA112797446	16.7 %
BOLD / iBOL	Barcode record	HTML landing page	http://portal.boldsystems.org/record/CHOLH011-12	16.7 %
BOLD / iBOL	Dataset	DOI	https://doi.org/10.5883/DP-Latest	16.7 %
LifeWatch ERIC	Metadata Catalogue	HTML landing page (JS app)	https://metadatalogue.lifewatch.eu/srv/eng/catalog.search#/metadata/dc512697-...	16.7 %
LifeWatch ERIC	Sample record	RDF/XML export	https://metadatalogue.lifewatch.eu/srv/api/records/dc512697-.../formatters/rdf?output=xml	16.7 %
eLTER	Dataset	DOI	https://doi.org/10.82159/gt66-zt98	16.7 %
DiSSCo	Digital specimen	DOI (landing page)	https://doi.org/10.3535/ZVR-CRA-1Y1	16.7 %
DiSSCo	Digital specimen	DOI metadata (? noredirect)	https://doi.org/10.3535/ZVR-CRA-1Y1?noredirect	16.7 %
DiSSCo	Digital specimen	REST API	https://disscover.dissco.eu/api/digital-specimen/v1/10.3535/ZVR-CRA-1Y1	16.7 %

However ...

LifeWatch ERIC	Same record	RDF/XML export	https://metadatalogue.lifewatch.eu/srv/api/records/dc512697-.../formatters/rdf?output=xml	16.7 %
ERGA	ERGA portal organism	HTML landing page	https://portal.erga-biodiversity.eu/organism/SAMEA112797446	16.7 %

16,7% is the minimal FAIR-checker score:

- web accessible resource
- no semantic metadata

malformed RDF/XML
no JSON-LD markup

EBI Biosamples	BioSamples structured data	JSON-LD export	https://www.ebi.ac.uk/biosamples/samples/SAMEA112797446.ldjson	45.8 %
----------------	----------------------------	----------------	---	--------

45% is not a good FAIR score for a well curated resource

- **Generic FAIR assessments may be useless for community needs**
- **Community-specific fields may not be evaluated by FAIR-checker**

Biodiv-FAIR-Checker project

Funding

ELIXIR Commissioned Service 2024-SCIENCE-BFSP, Work Package 6: raising the FAIRness of European biodiversity data through ELIXIR-aligned standards, services and training.

Communities

Consultations with ERGA, iBOL/BOLD, LifeWatch ERIC, eLTER, DiSSCo and GGBN: genomics, barcodes, specimens and ecological time series.

Partners

ELIXIR-FR (FAIR-Checker),
ELIXIR-CH (communities),
ELIXIR-UK (Bioschemas)

Objective #1

Identify community-specific practices in terms of metadata

Objective #2

Develop tooling through a community specific plugin in FAIR-Checker

Objective #3

Build and make available this profile through Bioschemas

BioDiv-FC: community-aware FAIR-Checker

Expected results:

- ▶ Extends the open-source **FAIR-Checker** tool with **plugins**
- ▶ 1 plugin:
 - Generic metrics kept, but **rephrased recommendations** for biodiversity
 - Community profile formalised in **SHACL** (mandatory, recommended, optional properties)
- ▶ A new **Biodiversity profile** extending the collection of Bioschemas profiles → improved visibility, community support, tooling ...

FAIR-Checker

<https://fair-checker.france-bioinformatique.fr>

FAIR-Checker consumes web pages

Resource identifier (URL/DOI)

 All metrics

Valid URL/DOI - The input contains the following DOIs that you can also test:  10.3205/09dgnc137

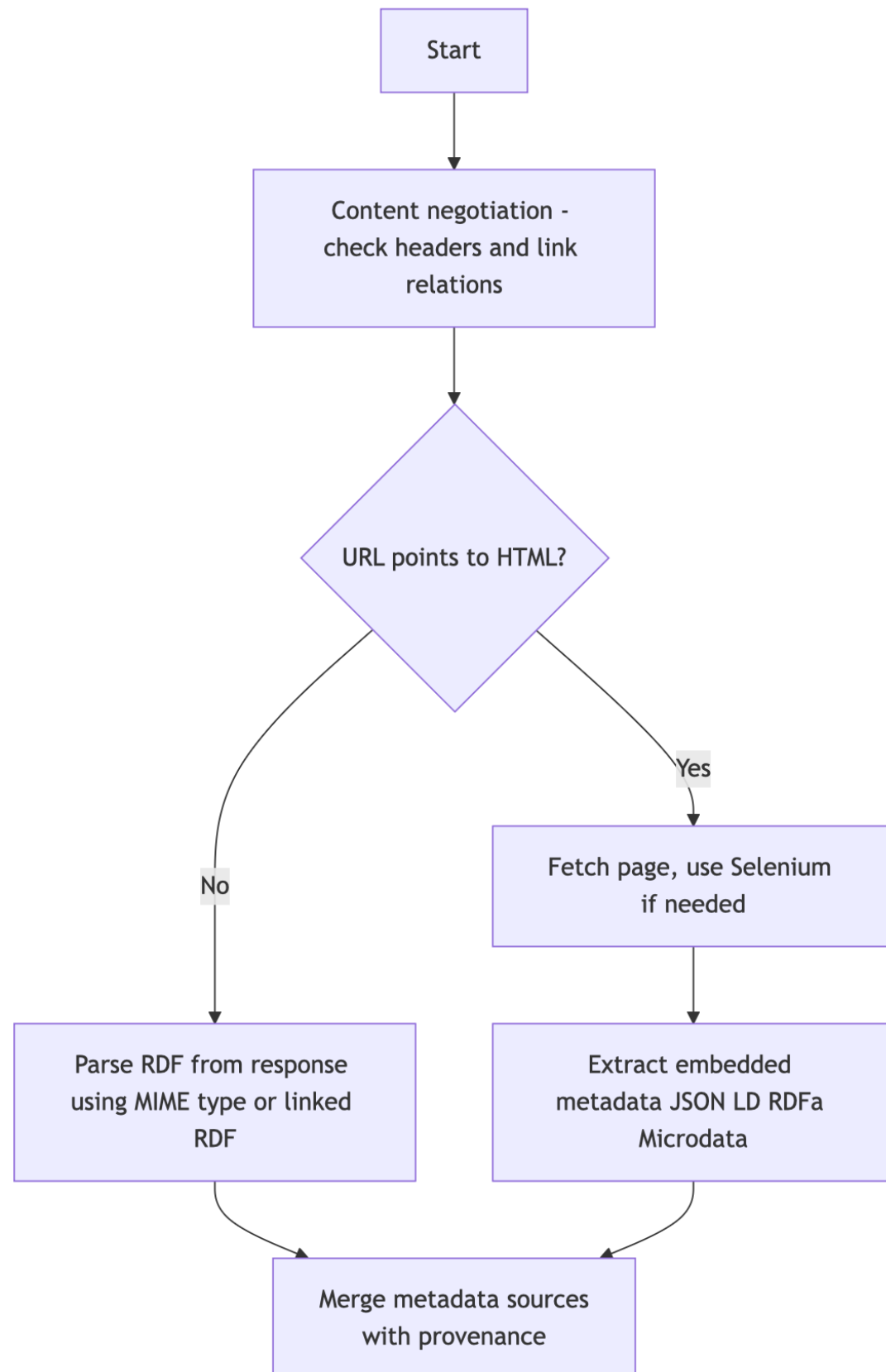


 Clean results

[Dataset Dataverse](#) [Workflow](#) [Publication Datacite](#) [Dataset](#) [Tool](#)

- ▶ Users submit web page **URLs** or **DOIs**.
- ▶ DOIs are resolved as web pages
- ▶ FAIR-Checker consumes the referred locations on the web

Advanced metadata harvesting



► **HTML rendering + parsing:**

Semantic metadata is extracted from the web page in JSON-LD, microdata, RDFa. (JSON-LD is the most adopted format)

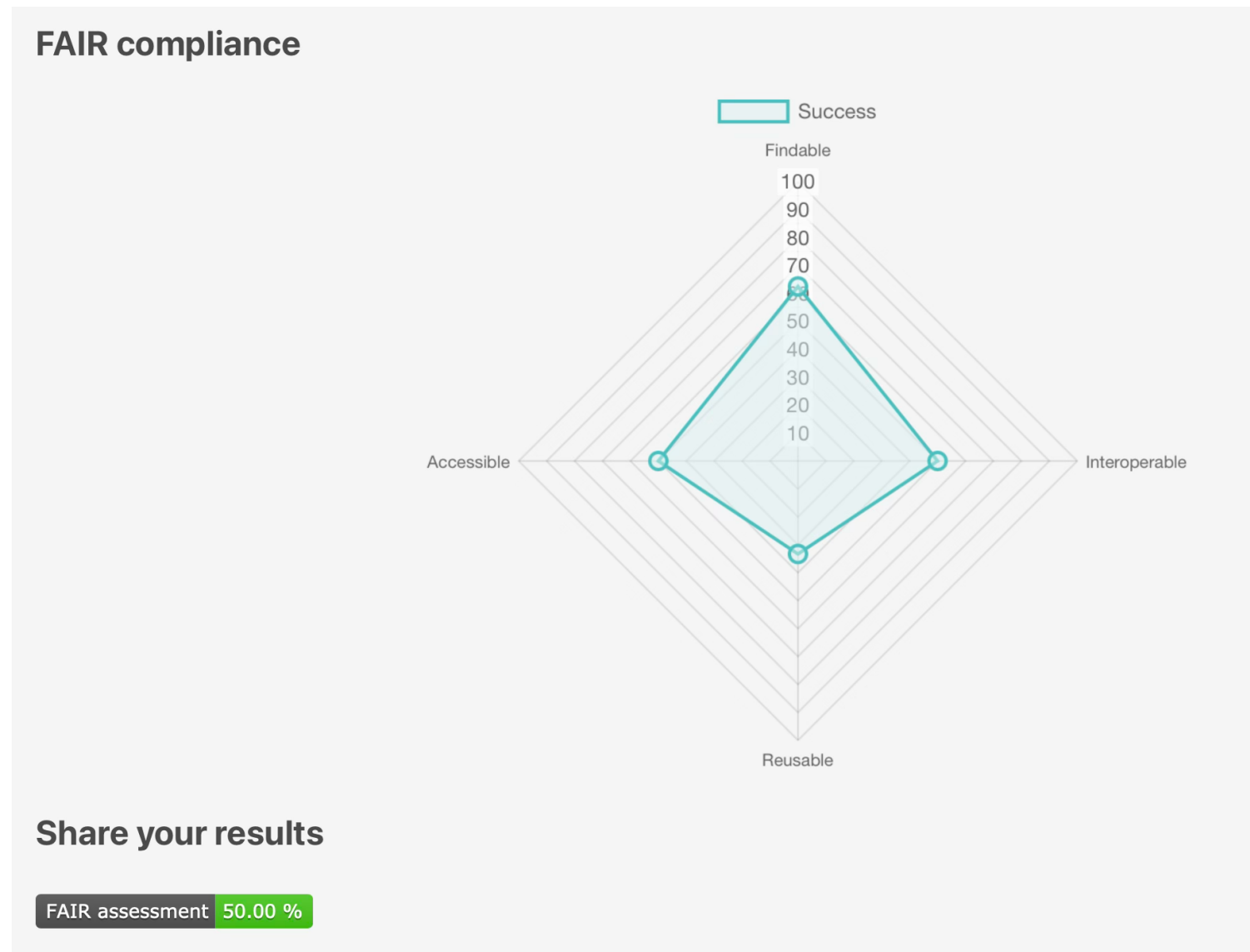
► **Content negotiation:**

Is the web server able to answer something different from a web page ? semantic metadata in RDF ?

► **FAIR Signposting:** a protocol to guide machine where the metadata is e.g. inside the web page, in a file on the server, at a remote location)

TODO

FAIR-Checker produces a FAIR assessment report



- ▶ Aggregated score per principle
- ▶ Visualised with a radar plot
- ▶ HTML badge summarising the whole evaluation

Detailed FAIR assessment results

F2B: Shared vocabularies for metadata	Check	FAIR principle F2B 2/2		i
A1.1: Open resolution protocol	Check	FAIR principle A1.1 2/2		i
A1.2: Authorisation procedure or access rights	Check	FAIR principle A1.2 0/2	You should describe the access policy in metadata by using at least one of the Read more ▾	i
I1: Machine readable format	Check	FAIR principle I1 1/2	You should provide discoverability oriented metadata with one of the following properties: dct:title dct:description dcat:accessURL dcat:downloadURL dcat:endpointDescription dcat:endpointURL Read less ▲	i
I2: Use shared ontologies	Check	FAIR principle I2 2/2		i
I3: External links	Check	FAIR principle I3 0/2	You should enrich your metadata with more diversified external links. Here we did not Read more ▾	i
R1.1: Metadata includes license	Check	FAIR principle R1.1 0/2	You should include information about licence in your metadata using one of the Read more ▾	i

- ▶ Per metric evaluation with $0 \leq s \leq 2$
- ▶ Recommendations for improvement if $s < 2$
- ▶ Detailed information on what is evaluated
- ▶ Metrics can be computed individually

FAIR-Checker (FC) evaluates semantic metadata

Findability F1B, F2	Accessibility A1.2	Reuse (licenses) R1.1	Reuse (provenance) R1.2
dct:identifier schema:identifier dct:title dct:description dcat:accessURL dcat:downloadURL dcat:endpointDescription dcat:endpointURL	odrl:hasPolicy dct:rights dct:accessRights	schema:license dct:license doap:license dbo:license cc:license xhv:license sto:license nie:license	prov:wasGeneratedBy prov:wasDerivedFrom prov:wasAttributedTo prov:used prov:wasInformedBy prov:wasAssociatedWith prov:startedAtTime prov:endedAtTime dct:hasVersion dct:isVersionOf dct:creator dct:contributor dct:publisher pav:hasVersion pav:version pav:hasCurrentVersion pav:createdBy pav:authoredBy pav:retrievedFrom pav:importedFrom pav:createdWith pav:retrievedBy pav:importedBy pav:curatedBy pav:createdAt pav:previousVersion schema:creator schema:author schema:publisher schema:provider schema:funder

"anyOf"

DC-Terms

DCAT

Schema.org

+ ODRL

+ DOAP,

+ PROV-O, PAV

DBO, CC ...

Table 2 Summary of the selected ontology properties relevant to assess three specific FAIR principles in *FAIR-Checker*

FC benefits from 37+ Life Science profiles



schema.org

- ▶ different usage of schema.org for life sciences
- ▶ Communities agree on **minimal**/
recommended/
optional annotation

Name	Group	Use Cases	Cross Walk	Task & Issues	Examples	Live Deploys
<u>ChemicalSubstance</u> (v0.4-RELEASE) 07 April 2020	<u>Chemicals</u>					
<u>ComputationalTool</u> (v1.0-RELEASE) 11 October 2021	<u>Tools</u>					
<u>ComputationalWorkflow</u> (v1.0-RELEASE) 09 March 2021	<u>Workflow</u>					
<u>DataCatalog</u> (v0.3-RELEASE-2019_07_01) 01 July 2019	<u>Data Repositories</u>					
<u>Dataset</u> (v0.3-RELEASE-2019_06_14) 14 June 2019	<u>Datasets</u>					
<u>FormalParameter</u> (v1.0-RELEASE) 09 March 2021	<u>Workflow</u>					
<u>Gene</u> (v1.0-RELEASE) 07 April 2021	<u>Genes</u>					
<u>MolecularEntity</u> (v0.5-RELEASE) 07 April 2020	<u>Chemicals</u>					
<u>Protein</u> (v0.11-RELEASE) 07 April 2020	<u>Proteins</u>					
<u>Sample</u> (v0.2-RELEASE-2018_11_10) 10 November 2018	<u>Samples</u>					
<u>Taxon</u> (v0.6-RELEASE) 07 April 2020	<u>Biodiversity</u>					

FC evaluate metadata completeness

R1.3: (Meta)data meet domain-relevant community standards

Validation of Bioschemas profiles:

- **rank missing** metadata
- developer **focus** on **minimal** metadata first

Check BioSchemas

`https://workflowhub.eu/workflows/18?version=1` has type `http://schema.org/ComputationalWorkflow`
Using `https://bioschemas.org/profiles/ComputationalWorkflow/1.0-RELEASE` for validation, specified from the **dct:conformsTo** property.

Required missing properties	Improvements
<code>https://schema.org/input</code> must be provided	<code>https://schema.org/citation</code> should be provided
<code>https://schema.org/output</code> must be provided	<code>https://schema.org/contributor</code> should be provided
	<code>https://schema.org/creativeWorkStatus</code> should be provided
	<code>https://schema.org/documentation</code> should be provided
	<code>https://schema.org/funding</code> should be provided
	<code>https://schema.org/hasPart</code> should be provided
	<code>https://schema.org/isBasedOn</code> should be provided
	<code>https://schema.org/maintainer</code> should be provided
	<code>https://schema.org/publisher</code> should be provided
	<code>https://schema.org/runtimePlatform</code> should be provided
	<code>https://schema.org/softwareRequirements</code> should be provided
	<code>https://schema.org/targetProduct</code> should be provided

ENA checklists as community
profiles for biodiversity
samples ?

Towards a Bioschemas "Biosample" profile

Checklist Fields

Filter fields...



Filter by type:

Marine Event

Part and
developmental stage
of organism

Organism
characteristics:
ecosystem

sample collection:
methods, storage
and transport

sample collection:
site related

sample collection:
core sample
properties

non-sample terms:
study or project

non-sample terms

Collection event
information

sample collection

Organism

Field Name		Field Format	(Field Restriction)	Requirement	(Units)
sample coordinator affiliation	?	free text		optional	
sample same as	?	restricted text	regular expression ?	optional	
sample derived from	?	restricted text	regular expression ?	optional	
project name	?	free text		mandatory	
barcoding center	?	free text		optional	
tolid	?	free text		optional	
collected_by	?	free text		mandatory	
collection date	?	restricted text	regular expression ?	mandatory	
geographic location (latitude)	?	restricted text	regular expression ?	recommended	DD
geographic location (longitude)	?	restricted text	regular expression ?	recommended	DD
geographic location (region and locality)	?	free text		mandatory	
identified_by	?	free text		optional	
elevation	?	restricted text	regular expression ?	optional	m

ENA
European Nucleotide Archive

Home Submit Search About Support

We recommend that you subscribe to the ENA-announce mailing list for updates c

European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the functional annotation. [More about ENA.](#)

3 priorities for metadata fields

Importance	Count	Elements
Mandatory	10	"organism part", "lifestage", "project name", "collected_by", "collection date", "geographic location (region and locality)", "habitat", "sex", "geographic location (country and/or sea)", "collecting institution",
Recommended	3	"geographic location (latitude)", "geographic location (longitude)", "specimen_voucher",
Optional	29	"Latitude Start", "Longitude Start", "Latitude End", "Longitude End", "relationship", "sample symbiont of", "symbiont", "sample collection method", "sample coordinator affiliation", "sample same as", "sample derived from", "barcoding center", "tolid", "identified_by", "elevation", "identifier_affiliation", "original collection date", "original geographic location", "original geographic location (latitude)", "original geographic location (longitude)", "sample coordinator", "GAL", "specimen_id", "GAL_sample_id", "proxy voucher", "proxy biomaterial", "bio_material", "culture_or_strain_id", "depth",

Each checklist attribute is a name/value pair

```
@prefix schema: <http://schema.org/> .

<https://www.ebi.ac.uk/biosamples/samples/xxx123456>
  rdf:type schema:Sample ;
  schema:additionalProperty [
    rdf:type schema:PropertyValue ;
    schema:name "habitat" ;
    schema:value "forest"
  ] , [
    schema:name "specimen_id" ;
    schema:value "xxx654321"
  ] , [
    schema:name "collection date" ;
    schema:value "missing" # present but not usable
  ] .
```

FAIR-Checker

implementation of the ENA53 checklist

Profile $\rightarrow$ graph shapes

$$P = \{C, M, R\}$$

$\rightarrow$ a metadata profile composed by target classes, mandatory and recommended properties

$C = \{C_1, C_2\} \rightarrow$ set of target classes to which the profile applies

$$M = \{p_1, p_2\}$$

$\rightarrow$ 2 mandatory properties

$$R = \{p_3, p_4, p_5\}$$

$\rightarrow$ 3 recommended properties

Generic SHACL template

```
ns:{{shape_name}} a sh:NodeShape ;
  {% for c in target_classes %}
    sh:targetClass {{c}} ;
  {% endfor %}

  {% for min_prop in min_props %}
    sh:property [
      sh:path {{min_prop}} ;
      sh:minCount 1 ;
      sh:severity sh:Violation
    ] ;
  {% endfor %}

  {% for rec_prop in rec_props %}
    sh:property [
      sh:path {{rec_prop}} ;
      sh:minCount 1 ;
      sh:severity sh:Warning
    ] ;
  {% endfor %}
```

...

Generated
SHACL constraints
for validating P



Shapes Constraint Language (SHACL)

W3C Recommendation 20 July 2017

This version:
<https://www.w3.org/TR/2017/REC-shacl-20170720/>

$\rightarrow$ 2 *sh:path* **strong** cardinality constraints on p_1 and p_2 and 3 **medium** cardinality constraints on p_3 , p_4 and p_5 for C_1 or C_2 instances.

Example SHACL sub-shape

We need to adjust the shape template to meet schema.org usage for ENA 53 checklist for *Biosamples*.

```
@prefix schema: <http://schema.org/> .

<https://www.ebi.ac.uk/biosamples/samples/xxx123456>
  rdf:type schema:Sample ;
  schema:additionalProperty [
    rdf:type schema:PropertyValue ;
    schema:name "habitat" ;
    schema:value "forest"
  ] , [
    schema:name "specimen_id" ;
    schema:value "xxx654321"
  ] , [
    schema:name "collection date" ;
    schema:value "missing" # present but not usable
  ] .
```

```
sh:property [
  sh:path schema:additionalProperty ;
  sh:qualifiedValueShape [
    sh:property [ sh:path schema:name ; sh:hasValue "habitat" ] ;
    sh:property [ sh:path schema:value ; sh:not [ sh:in ( "missing" "None" ) ] ]
  ] ;
  sh:qualifiedMinCount 1 ;
  sh:severity sh:Violation ; # Warning for recommended, Info for optional
] ;
```

Listing 4: part of a generated SHACL shape aimed at validating the presence of at least one “habitat” metadata field, without “missing” or “None” values.

Metadata completeness score

Each level is assigned an importance weight **3 for mandatory**, **2 for recommended** and **1 for optional** and the completeness score is the fraction of achievable weighted points, expressed as a percentage.

$$score = \frac{3(M - e) + 2(R - w) + (O - i)}{3M + 2R + O} \times 100$$

M, R, O : number of mandatory, recommended, and optional properties

e, w, i : number of SHACL shape "errors", "warning", "errors"

Demo

<https://www.ebi.ac.uk/biosamples/samples/SAMEA119944482>

FAIR  checker by 

CheckInspectAbout usAPI

DocumentationStatisticsFeedbacksGithub

Step 1: fetch RDF metadata from the resource URL

Examples: [Dataset Dataverse](#) [Workflow](#) [Publication Datacite](#) [Dataset](#) [Tool](#)



Build Knowledge Graph

Retrieved Knowledge Graph

... from Wikidata

... from OpenCitations

... from OpenAIRE

```
@prefix sc: <http://schema.org/> .

<https://www.ebi.ac.uk/biosamples/samples/SAMEA119944482#mime-probe> {
  <http://identifiers.org/biosample/SAMEA119944482> a sc:DataRecord ;
  sc:dateCreated "2025-08-26T14:55:47.745Z" ;
  sc:dateModified "2025-08-26T17:00:05.34Z" ;
  sc:dateReleased "2025-08-26T17:00:04.602Z" ;
  sc:dateSubmitted "2025-08-26T14:55:47.745Z" ;
  sc:identifier "biosample:SAMEA119944482" ;
  sc:isPartOf <https://www.ebi.ac.uk/biosamples/samples> ;
  sc:mainEntity <https://www.ebi.ac.uk/biosamples/samples/SAMEA119944482> .

<https://www.ebi.ac.uk/biosamples/samples> a sc:Dataset .

<https://www.ebi.ac.uk/biosamples/samples/SAMEA119944482> a <OBI:0000747>,
  sc:Sample .
```

https://www.ebi.ac.uk/biosamples/samples/SAMEA119944482#mime-probe: 136 Triples

JSON-LD

Trig

Step 2: Metadata quality checks

 Controlled vocabularies

 Bioschemas

 Datacite

 ENA53

Demo

<https://www.ebi.ac.uk/biosamples/samples/SAMEA119944482>

Check ENA53 profile

EVALUATED RESOURCE

<https://www.ebi.ac.uk/biosamples/samples/SAMEA119944482>

COMPLETENESS SCORE

67.69%

Weighted score: required properties count 3x, recommended 2x, optional 1x — measuring how many ENA Checklist 53 properties are actually provided.

Required missing properties

Recommended missing properties

Optional missing properties

Latitude Start could be provided

Longitude Start could be provided

Latitude End could be provided

Longitude End could be provided

relationship could be provided

sample symbiont of could be provided

sample collection method could be provided

sample coordinator affiliation could be provided



Now, moving from a non-semantic checklist to Bioschemas ?

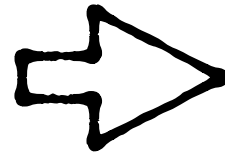
schema.org



Bioschemas

Work in progress: schema.org alignment of key-value pairs

Importance	Count	Elements
Mandatory	10	"organism part", "lifestage", "project name", "collected_by", "collection date", "geographic location (region and locality)", "habitat", "sex", "geographic location (country and/or sea)", "collecting institution",
Recommended	3	"geographic location (latitude)", "geographic location (longitude)", "specimen_voucher",
Optional	29	"Latitude Start", "Longitude Start", "Latitude End", "Longitude End", "relationship", "sample symbiont of", "symbiont", "sample collection method", "sample coordinator affiliation", "sample same as", "sample derived from", "barcoding center", "tolid", "identified_by", "elevation", "identifier_affiliation", "original collection date", "original geographic location", "original geographic location (latitude)", "original geographic location (longitude)", "sample coordinator", "GAL", "specimen_id", "GAL_sample_id", "proxy voucher", "proxy biomaterial", "bio_material", "culture_or_strain_id", "depth",



Property	Expected Type	Description
Properties from BioSample (pending schema.org integration).		
scientificName	DefinedTerm Text TaxonName	The full scientific name (order, family, subfamily, genus, specific epithet).
anatomicalSystem	AnatomicalSystem or DefinedTerm URL	Indicate the anatomical system from which the sample was collected, e.g. circulatory, digestive, endocrine, immune and other systems.
anatomicalStructure	AnatomicalStructure or DefinedTerm URL	Indicate the organ or tissues the sample was collected from.
sex	Text DefinedTerm URL	The biological sex of the individual from which the sample was obtained.
developmentalStage	Text QuantitativeValue DefinedTerm	A temporal subdivision of a developmental process, indicating the age or stage of the object when the sample was created.
associatedPhenotype	Text QuantitativeValue DefinedTerm bioschemastypesdrafts:Phenotype	The relevant defining characteristic(s) expressed by the biological source of a typical sample, usually physical.
cellType	Text URL DefinedTerm	The cell line, lineage or biological/taxonomic classification of a cell.
Properties from BioChemEntity		
associatedDisease	MedicalCondition or PropertyValue or URL	Disease associated to this BioChemEntity. Such a disease can be a MedicalCondition or a URL. If you want to add an evidence supporting the association, please use PropertyValue.
biologicalRole	DefinedTerm	A role played by the molecular entity within a biological context.
isInvolvedInBiologicalProcess	DefinedTerm or PropertyValue	Biological process this BioChemEntity is involved in; please use PropertyValue if you want to



- ▶ Can be **time** consuming
- ▶ Requires **domain** expertise
- ▶ Requires **schema.org** + semantic web expertise

Work in progress: schema.org alignment of key-value pairs

1. Use an **LLM for drafting** the semantic alignment
 - Provide the key-value pairs (checklist)
 - Provide the Biosample type
 - Explain what is and how is written an SSSOM mapping
 - Ask for confidence scores
2. **Manually revise** the mappings with experts
 - focus on low confidence scores

Work in progress: schema.org alignment of key-value pairs

15	subject_id	subject_label	predicate_id	object_id	object_label	mapping_justification	author_id	confidence	mapping_tool
16	# --- Mandatory elements ---								
17	ENA_53:organism_part	<u>organism part</u>	skos:closeMatch	bioschemas:anatomicalStructure	anatomicalStructure	semapv:ManualMappingCuration		0.75	claude-sonnet-4-6
18	ENA_53:lifestage	lifestage	skos:closeMatch	bioschemas:developmentalStage	developmentalStage	semapv:ManualMappingCuration		0.85	claude-sonnet-4-6
19	ENA_53:project_name	project name	skos:closeMatch	bioschemas:studyTitle	studyTitle	semapv:ManualMappingCuration		0.65	claude-sonnet-4-6
20	ENA_53:collected_by	collected_by	skos:exactMatch	bioschemas:collector	collector	semapv:ManualMappingCuration		0.95	claude-sonnet-4-6
21	ENA_53:collection_date	collection date	skos:exactMatch	bioschemas:dateCollected	dateCollected	semapv:ManualMappingCuration		0.95	claude-sonnet-4-6
22	ENA_53:geographic_location_region_and_location	geographic location (region and location)	skos:closeMatch	bioschemas:locationOfOrigin	locationOfOrigin	semapv:ManualMappingCuration		0.75	claude-sonnet-4-6
23	ENA_53:habitat	habitat	skos:closeMatch	bioschemas:environmentalSystem	environmentalSystem	semapv:ManualMappingCuration		0.80	claude-sonnet-4-6
24	ENA_53:sex	sex	skos:exactMatch	bioschemas:sex	sex	semapv:ManualMappingCuration		0.99	claude-sonnet-4-6
25	ENA_53:geographic_location_country_and_or_location	geographic location (country and/or location)	skos:closeMatch	bioschemas:itemLocation	itemLocation	semapv:ManualMappingCuration		0.78	claude-sonnet-4-6
26	ENA_53:collecting_institution	collecting institution	skos:closeMatch	bioschemas:collector	collector	semapv:ManualMappingCuration		0.65	claude-sonnet-4-6
27									

15	subject_id	subject_label	predicate_id	object_id	object_label	mapping_justification	author_id	confidence	mapping_tool	comment	Manual
50	ENA_53:original_collection_date	original collection date	skos:closeMatch	bioschemas:dateCollected	dateCollected	semapv:ManualMappingCuration		0.55	claude-sonnet-4-6	The original collection date refers to an earlier or amended collection date; dateCollected is the best available match but does not distinguish original vs revised dates.	
51	ENA_53:original_geographic_location	original geographic location	skos:closeMatch	bioschemas:locationOfOrigin	locationOfOrigin	semapv:ManualMappingCuration		0.55	claude-sonnet-4-6	The original geographic location corresponds to locationOfOrigin but implies the value may have been subsequently revised.	
52	ENA_53:original_geographic_location_latitude	original geographic location (latitude)	skos:broadMatch	bioschemas:locationOfOrigin	locationOfOrigin	semapv:ManualMappingCuration		0.42	claude-sonnet-4-6	Original latitude before possible revision; partially covered by locationOfOrigin via GeoCoordinates.	
53	ENA_53:original_geographic_location_longitude	original geographic location (longitude)	skos:broadMatch	bioschemas:locationOfOrigin	locationOfOrigin	semapv:ManualMappingCuration		0.42	claude-sonnet-4-6	Original longitude before possible revision; partially covered by locationOfOrigin via GeoCoordinates.	
54	ENA_53:sample_coordinator	sample coordinator	skos:closeMatch	schema:contactPoint	contactPoint	semapv:ManualMappingCuration		0.70	claude-sonnet-4-6	The sample coordinator is a point of contact for the sample; maps to schema:contactPoint (Person or Organization).	
55	ENA_53:GAL	GAL			collector	semapv:ManualMappingCuration			claude-sonnet-4-6	No mapping: Genome Acquisition Lab name; an infrastructure/organisational field with no counterpart in Bioschemas BioSample.	
56	ENA_53:specimen_id	specimen_id	skos:closeMatch	bioschemas:alternateIdentifier	alternateIdentifier	semapv:ManualMappingCuration		0.75	claude-sonnet-4-6	A specimen identifier is an alternate identifier for the sample instance.	
57	ENA_53:GAL_sample_id	GAL_sample_id	skos:closeMatch	bioschemas:alternateIdentifier	alternateIdentifier	semapv:ManualMappingCuration		0.60	claude-sonnet-4-6	The Genome Acquisition Lab sample ID is a local alternate identifier for the sample.	
58	ENA_53:proxy_voucher	proxy voucher			schema:relatedTo	semapv:ManualMappingCuration			claude-sonnet-4-6	No mapping: a voucher specimen used as proxy for the sequenced material; the physical archiving and proxy relationship are not modelled in Bioschemas BioSample.	
59	ENA_53:proxy_biomaterial	proxy biomaterial			schema:relatedTo	semapv:ManualMappingCuration			claude-sonnet-4-6	No mapping: biomaterial used as a proxy for the primary sample; the proxy derivation chain is not captured in Bioschemas BioSample.	
60	ENA_53:bio_material	bio_material			isPartOfBioChemEntity	semapv:ManualMappingCuration			claude-sonnet-4-6	No mapping: identifier for a biological material held in a collection or biobank; Bioschemas BioSample lacks a dedicated property for biobank accession references.	
										A culture or strain identifier is a domain-specific alternate identifier	

Conclusion and next steps

Take-home messages

A very short project (1 year funding)

Already multiple **achievements**

- Community needs
- Prototype based on the ENA 53 checklist
- A draft Biosample profile

Next steps

- A plugin architecture to extend FAIR-Checker with community specific rules
- A Biosample profile within Bioschemas
- A roadmap for the development of other community-based plugins

The team



Swiss Institute of
Bioinformatics



INSTITUT FRANÇAIS DE BIOINFORMATIQUE



The University of Manchester

Coordination, community needs

Robert Waterhouse

Chiara Bortoluzzi

Christian de Guttery

Tooling

Alban Gaignard

Fred de Lamotte

Brieuc Quemeneur

Erwan Corre

Community engagement

Nick Juty

Carole Goble

Backup slides

FAIR assessment badges

FAIR assessment 66.67 %

```
@prefix : <https://fair-checker.france-bioinformatique.fr/data/> .
@prefix dqv: <http://www.w3.org/ns/dqv#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

:696eadd35f87f91deea979e4 a dqv:QualityMeasurement ;
  rdfs:seeAlso <https://doi.org/10.1186/s13326-023-00289-5> ;
  dqv:computedOn
<https://api.datacite.org/application/vnd.schemaorg.ld+json/10.7892/boris.108387> ;
  dqv:value "66.67"^^xsd:integer ;
  prov:generatedAtTime "2026-01-19T22:18:59.908000"^^xsd:dateTime ;
  prov:wasAttributedTo <https://github.com/IFB-ElixirFr/fair-checker> ;
  prov:wasDerivedFrom :696eadca5f87f91deea979cc,
    :696eadca5f87f91deea979d5,
    :696eadca5f87f91deea979d6,
    :696eadca5f87f91deea979d7,
    :696eadca5f87f91deea979d8,
    :696eadca5f87f91deea979d9,
    :696eadca5f87f91deea979da,
    :696eadca5f87f91deea979db,
    :696eadca5f87f91deea979dc,
    :696eadd35f87f91deea979de,
    :696eadd35f87f91deea979e0,
    :696eadd35f87f91deea979e2 .
```

- ▶ A badge points to a **persistent**, **machine-readable** result
- ▶ The evaluation result is FAIR itself (DQV, PROV ontologies):
 - ▶ **typed** entities
 - ▶ **linked** to individual metrics and the evaluated resource
 - ▶ with provenance metadata (*wasDerivedFrom*, *wasAttributedTo*)

Used LLM prompts

1 From this web page <https://bioschemas.org/types/BioSample/0.2-DRAFT>, can you extract the semantic properties and their related classes into a JSON data structure? Can you store this file in the project?

2 Here is a list of ENA metadata for the checklist numbered 53:

```
mandatory_elements = [  
    "organism part",  
    "lifestage",  
    "project name",  
    ...  
    "culture_or_strain_id",  
    "depth",  
]
```

Can you store them in a JSON data structure?

3 Very good, now this page <https://mapping-commons.github.io/sssom/dev/> describes SSSOM, a tabular format aimed at describing mappings between ontology classes and properties. Here are the column titles for such a terminology mapping format: ["subject_id", "subject_label", "predicate_id", "object_id", "object_label", "mapping_justification", "author_id", "confidence", "comment"] Can you generate an SSSOM mapping based on all ENA_53_checklist.json as subject_id and the Bioschemas Biosample properties you just extracted? Please report your confidence score for each mapping. The "subject_id" column should contain ENA metadata fields, the "object_id" should contain the Biosample properties. Store the SSOM mapping in a CSV file. Do not invent mapping when they do not exist.

4 can you replace the mapping_justification by "LanguageBasedSimilarityMeasure" to reflect that it was computed by an LLM? is there any other term for that?

5 for each of the 12 unmapped properties, can you propose schema.org candidate properties?