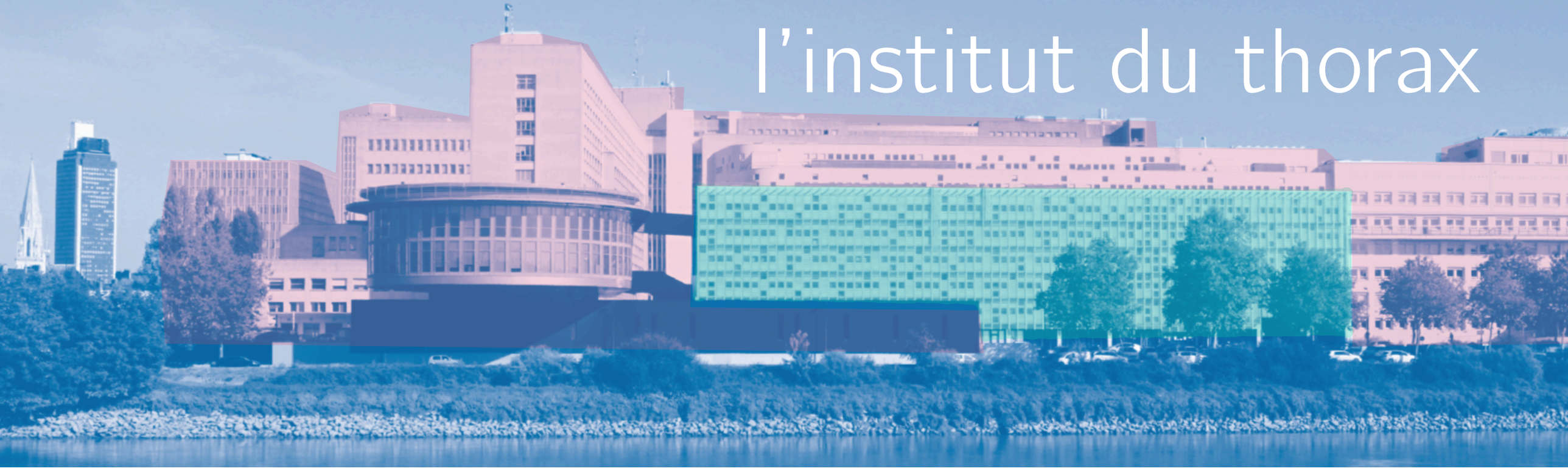
FAIR assements tailored to biodiversity resources ?

Alban Gaignard
CNRS, institut du thorax, Nantes, France

Jannuary 21, 2026

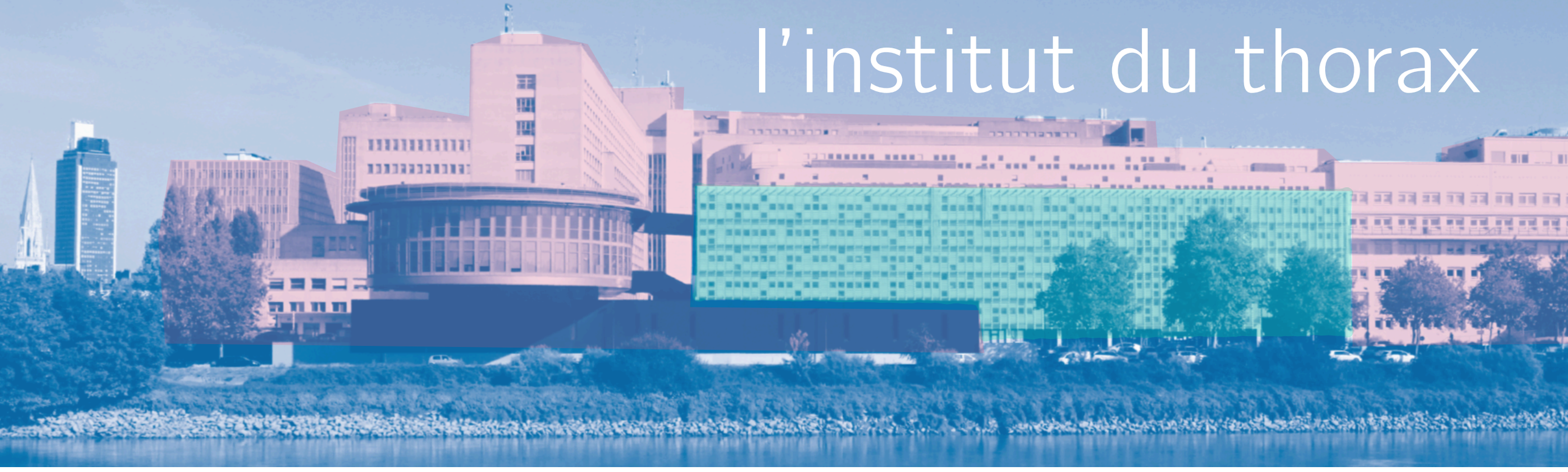
BioDiv-FAIRChecker kick-off meeting
SIB, Lausanne, Switzerland

Who am I?



l'institut du thorax

l'institut du thorax



l'institut du thorax

Better understanding of
cardio-vascular and
metabolic diseases

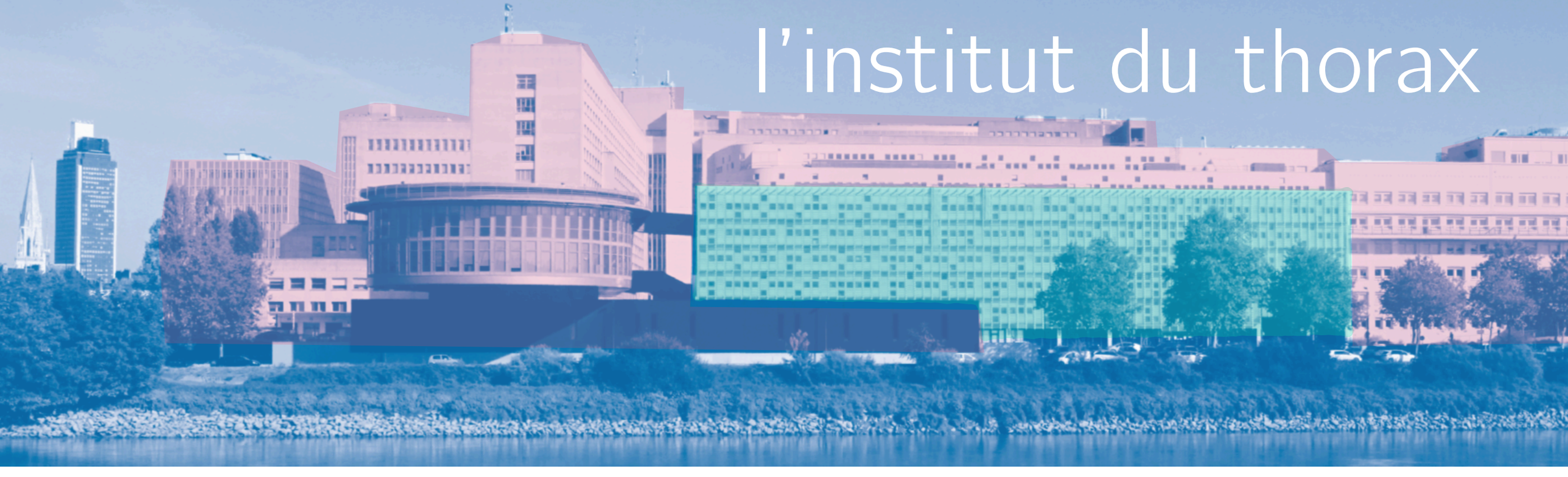
Gene ←→ function

associations

Translational medicine

university hospital ⊕ reseach lab

l'institut du thorax



l'institut du thorax

Better understanding of cardio-vascular and metabolic diseases

Gene ←→ function

associations

Translational medicine

university hospital ⊕ reseach lab

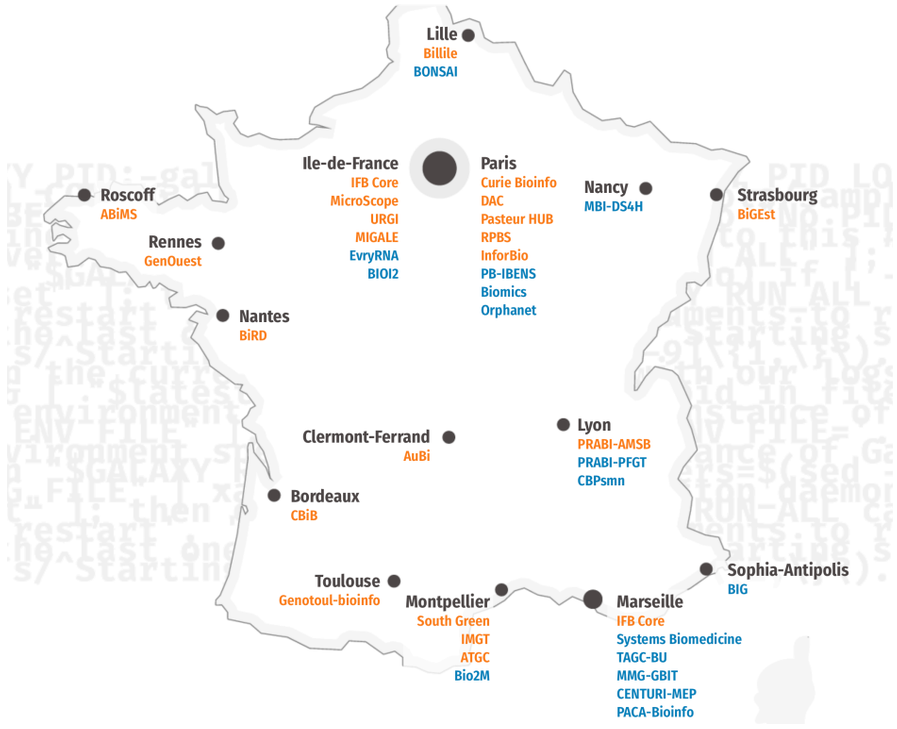
Bioinformatics

- Massive production of genomic sequence & health data
  → Workflows + HPC
- Integration of multi-modal and multi-scale data
- Predictive models

l'institut du thorax

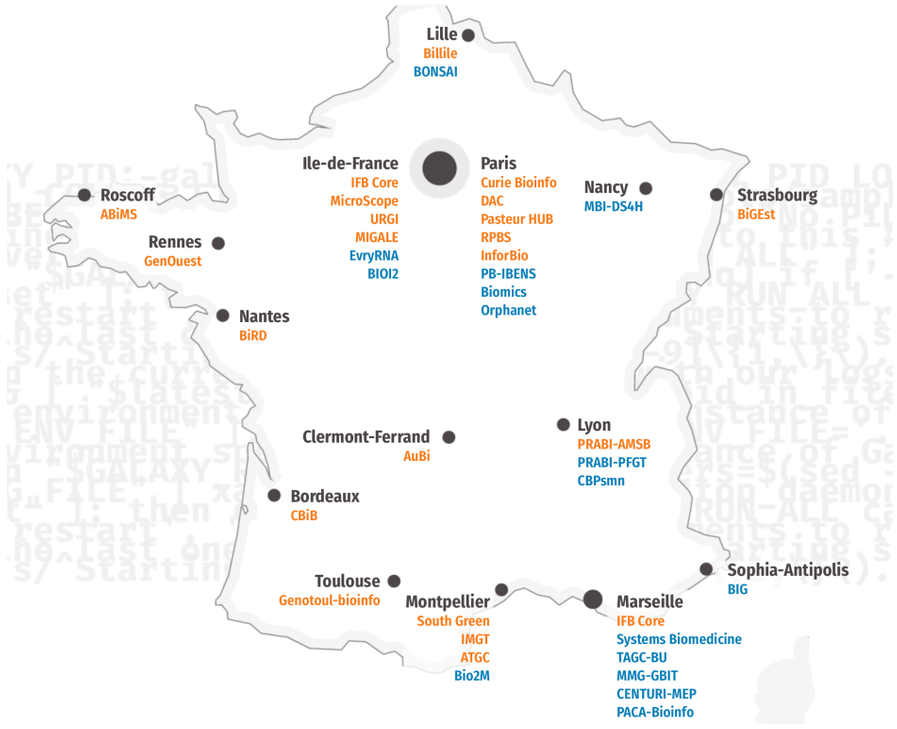
IFB = Elixir-FR

INSTITUT FRANÇAIS DE BIOINFORMATIQUE



IFB = Elixir-FR

INSTITUT FRANÇAIS DE BIOINFORMATIQUE



A national research infrastructure
for Bioinformatics providing:

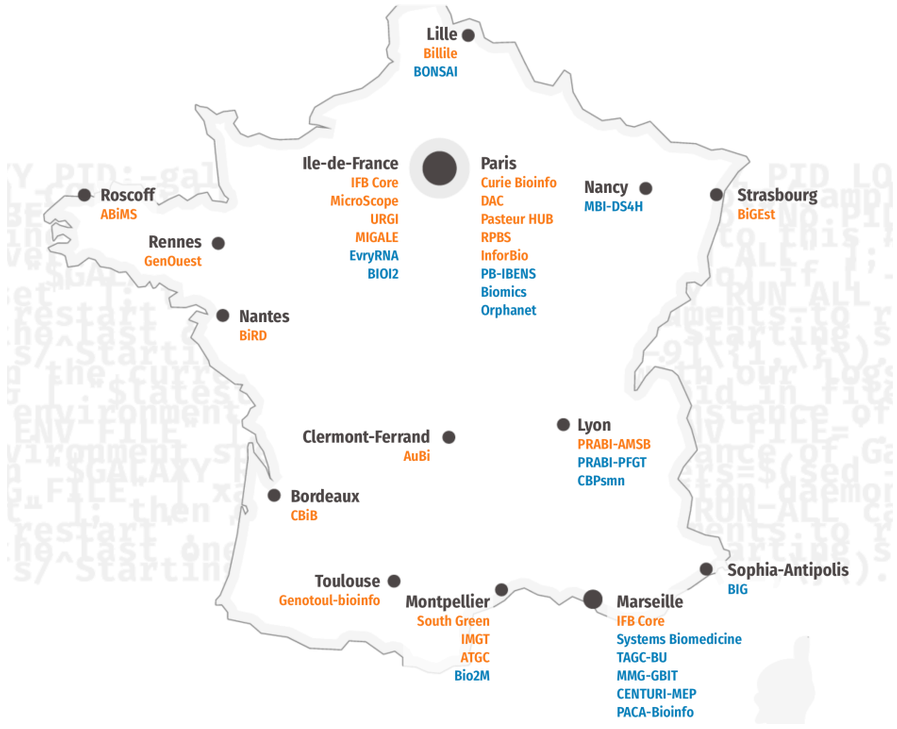
Compute & Storage

Tools & Workfows, Databases,
Training, Open Sciences

Communities: health, agronomy,
biodiversity, microbiology

IFB = Elixir-FR

INSTITUT FRANÇAIS DE BIOINFORMATIQUE



A national research infrastructure for Bioinformatics providing:

Compute & Storage

Tools & Workfows, Databases, Training, Open Sciences

Communities: health, agronomy, biodiversity, microbiology

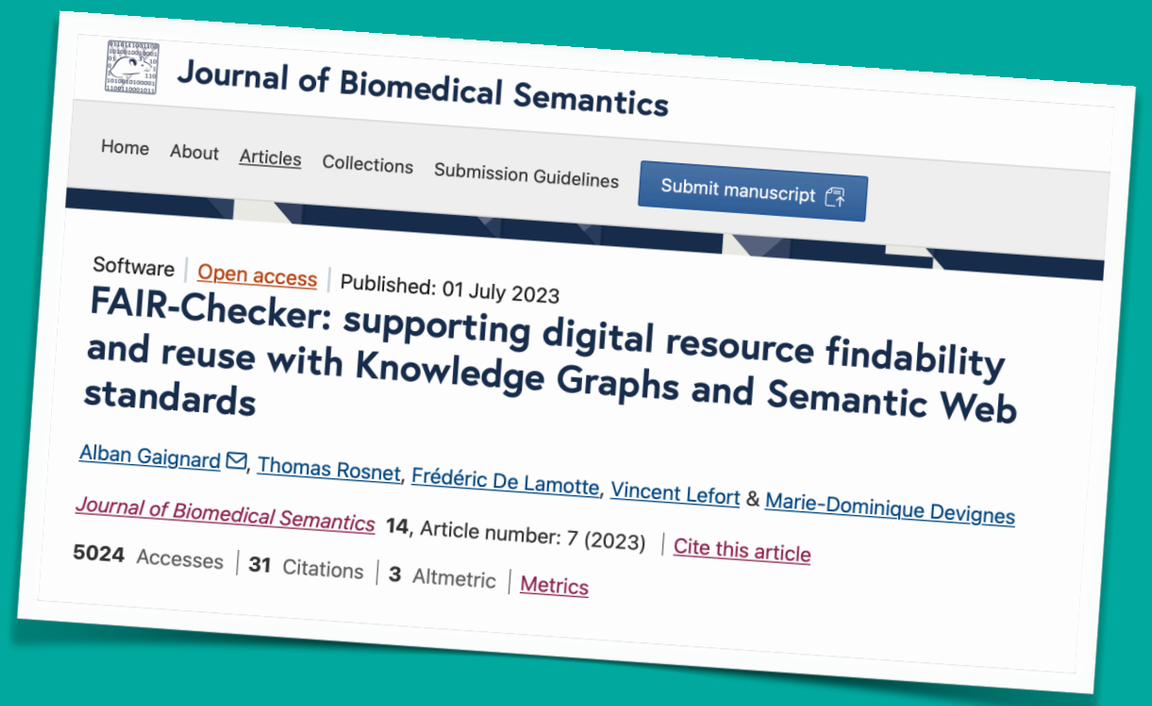
At IFB, I'm co-leading

- **Open sciences & interoperability**
  FAIR-Checker, metadata standards, ontologies (Bioschemas, EDAM), data management plans
- **Health community**:
  genomic data discoverability & sharing (Beacon, FEGA) + data integration (Knowledge Graphs)

Thomas Rosnet

Marie-Dominique Devignes

Frédéric de Lamotte



How to evaluate research data FAIRness ?

[1] Gaignard, A., Rosnet, T., de Lamotte, F., Lefort, V., & Devignes, M. (2023). *FAIR-Checker: supporting digital resource findability and reuse with Knowledge Graphs and Semantic Web standards*. **Journal of Biomedical Semantics**, 14. https://doi.org/10.1186/s13326-023-00289-5

FAIR principles require tooling

Australian Research Data Commons

FAIR principles

- critical for open & reproducible sciences
- result in many guidelines
- technology agnostic guidelines

How to implement the principles ...

... and go beyond checklists ?

Resource provider

and developers

need help and tooling.

https://www.go-fair.org/fair-principles

https://www.nature.com/articles/sdata201618

FAIR principles require tooling

Australian Research Data Commons

Web URIs *HTTP, RDF/SPARQL*

Domain ontologies

FAIR principles

- critical for open & reproducible sciences
- result in many guidelines
- technology agnostic guidelines

How to implement the principles ...

... and go beyond checklists ?

Resource provider

and developers

need help and tooling.

https://www.go-fair.org/fair-principles

https://www.nature.com/articles/sdata201618

Usage scenarios

- Datasets
- Training
- Tools
- etc ...

- Dataverse
- Bio.tools
- Zenodo,
  Pubmed ...

Where to publish ?
Which registry ?
Does it provide metadata ?
Is it enough to be FAIR ?

Improve metadata quality ?
Community specific standards ?
Which technology ?

Why a (nother) tool ?

Assumptions

- **"Linked Data" and Semantic Web technologies are key in most of the FAIR principles (especially F, I, and R)**

  ... but technical skills are needed.

Objectives

- Provide a web interface for resource providers to **evaluate FAIR metrics** and **make progress on FAIRification** (iterative testing)
- Provide additional tools for developers Leverage semantic web technologies (RDF, SPARQL, SHACL) to **enhance the quality of metadata**

FAIR Checker

General approach

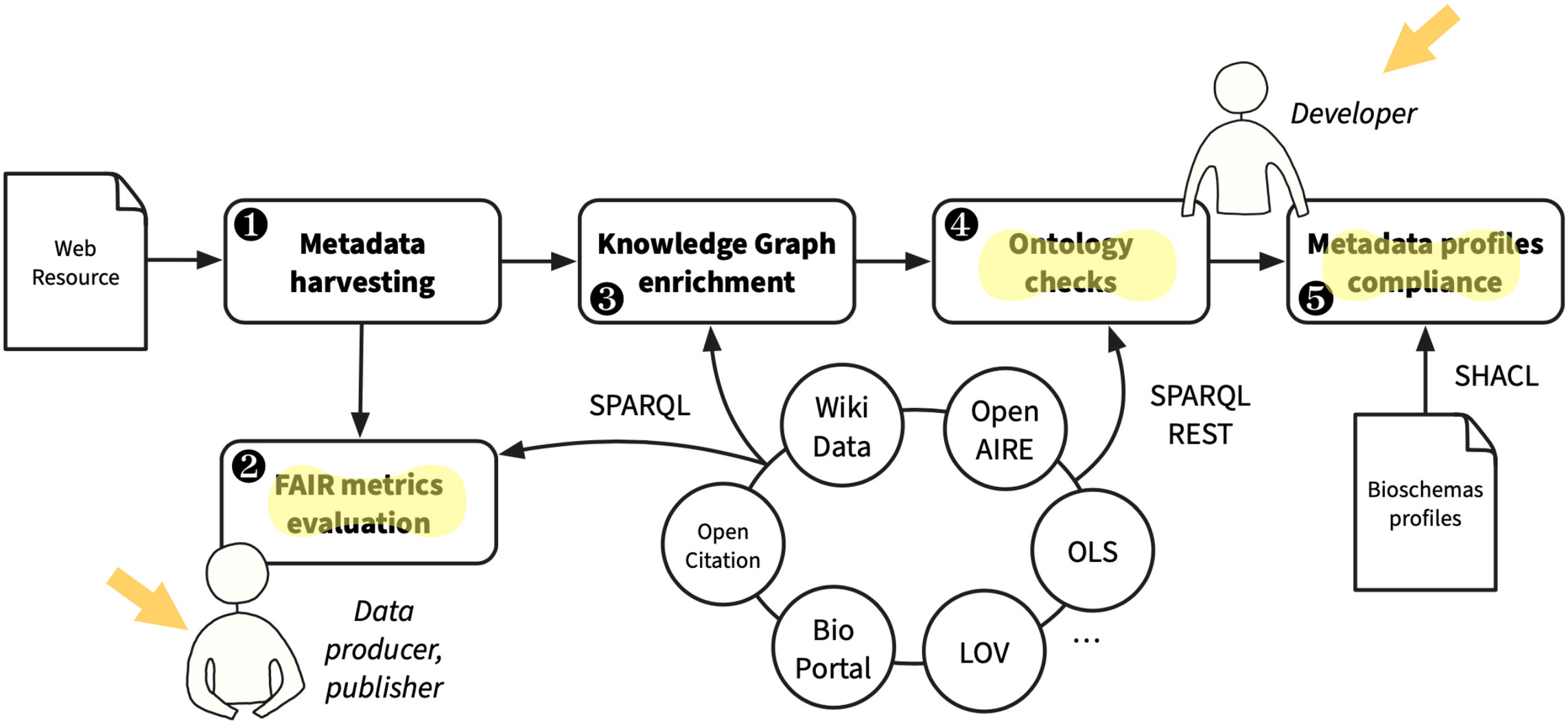
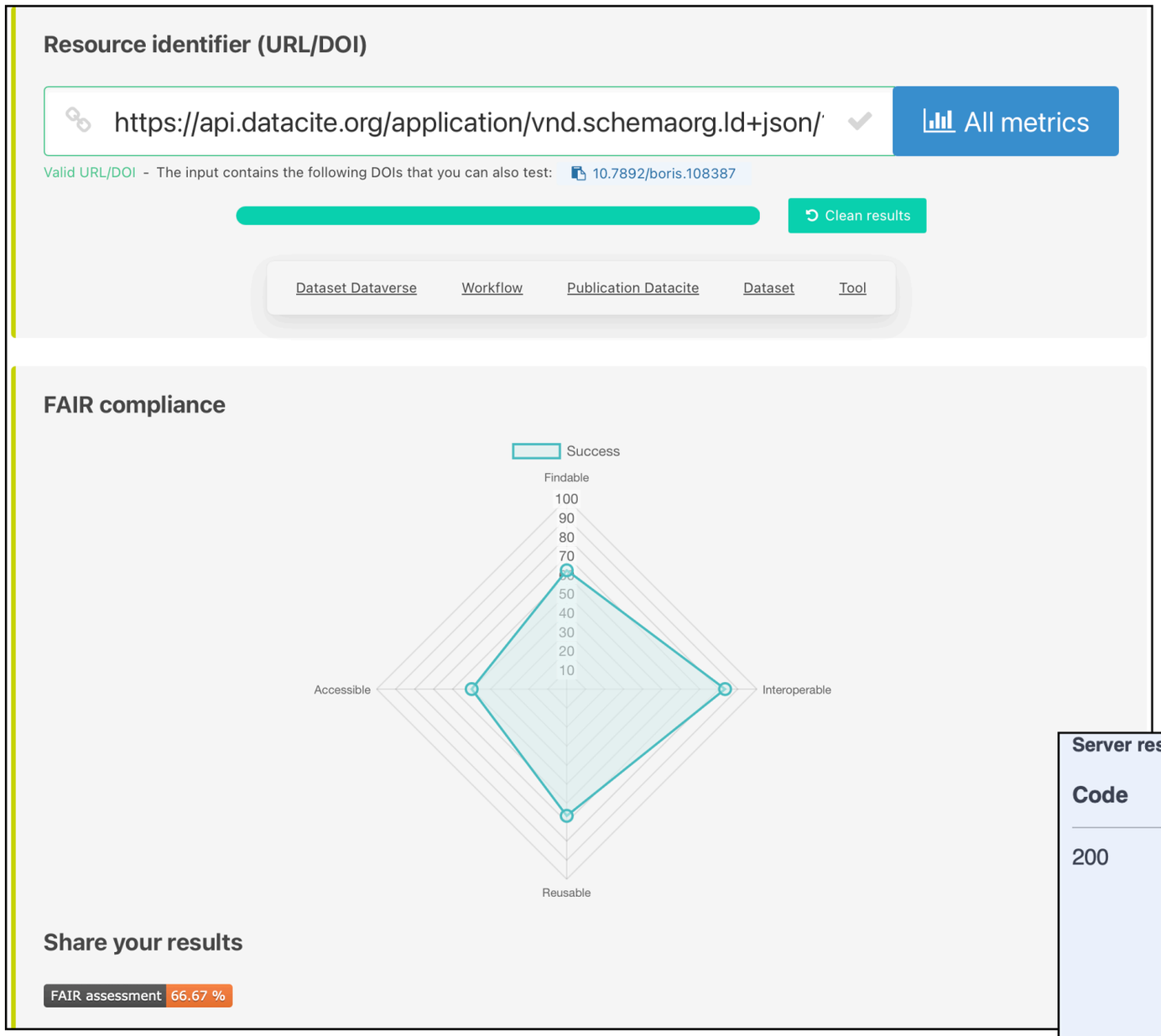
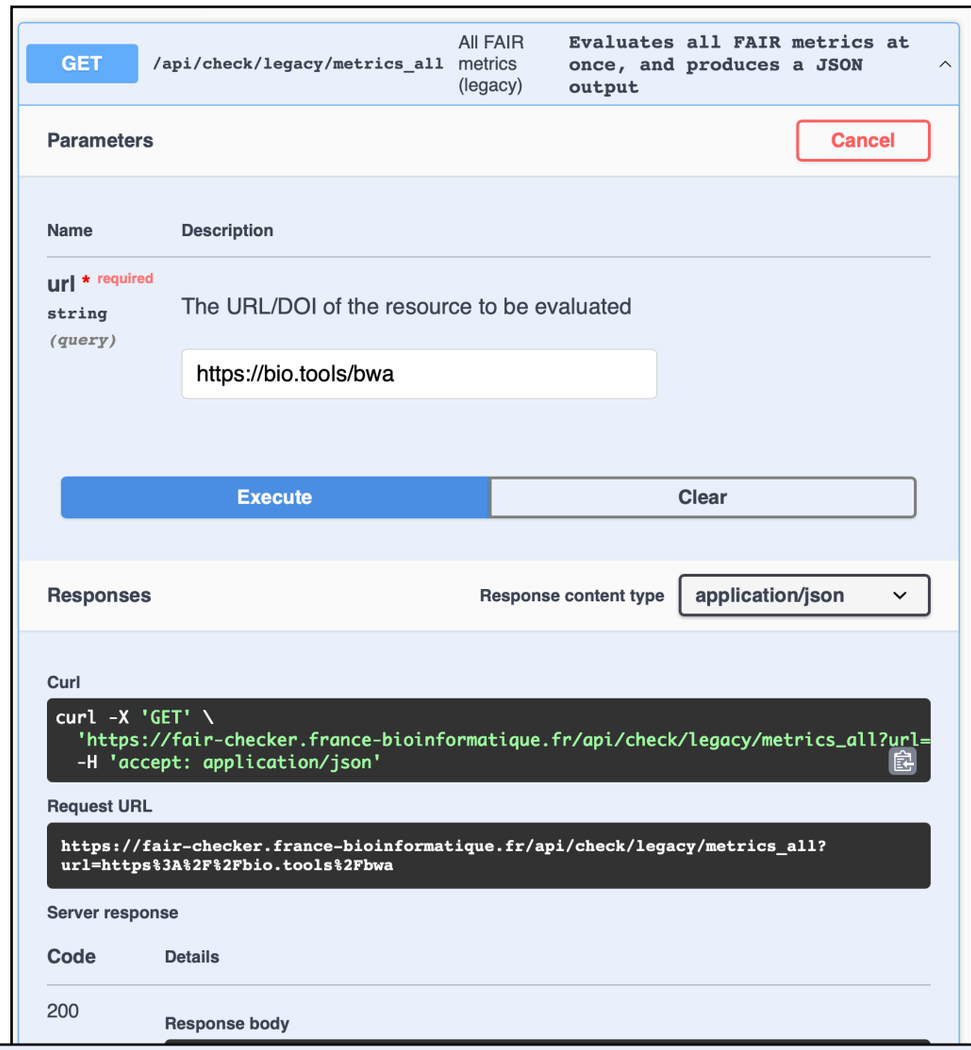


Figure 1 Gathering, enriching and analyzing semantic web annotations in line with FAIR principles.

A web UI + an API

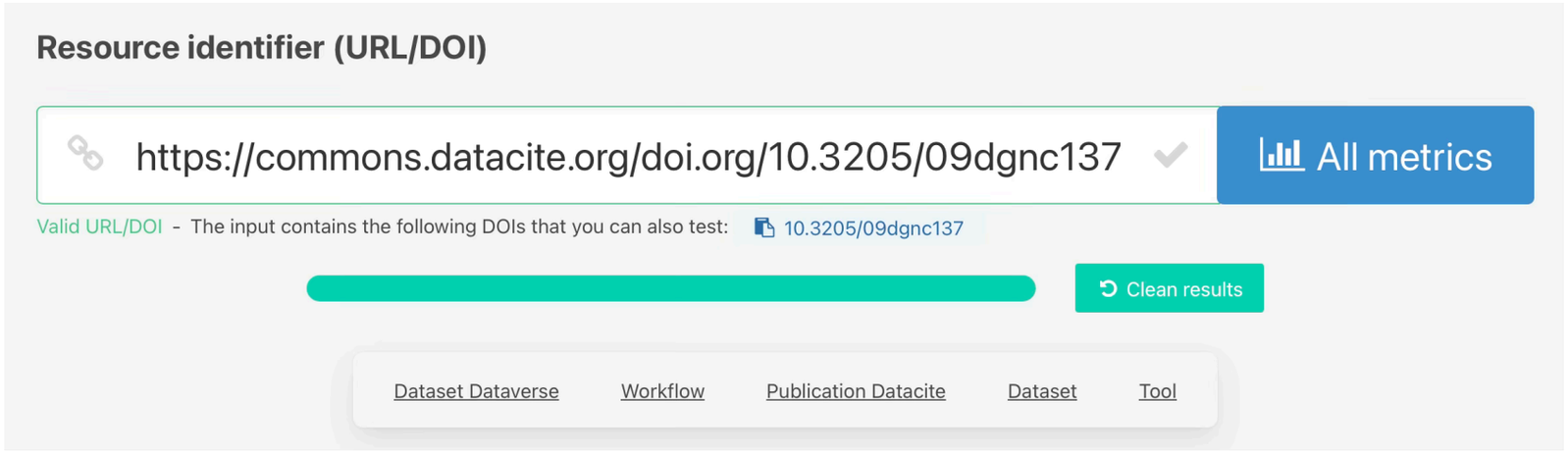




FAIR-Checker

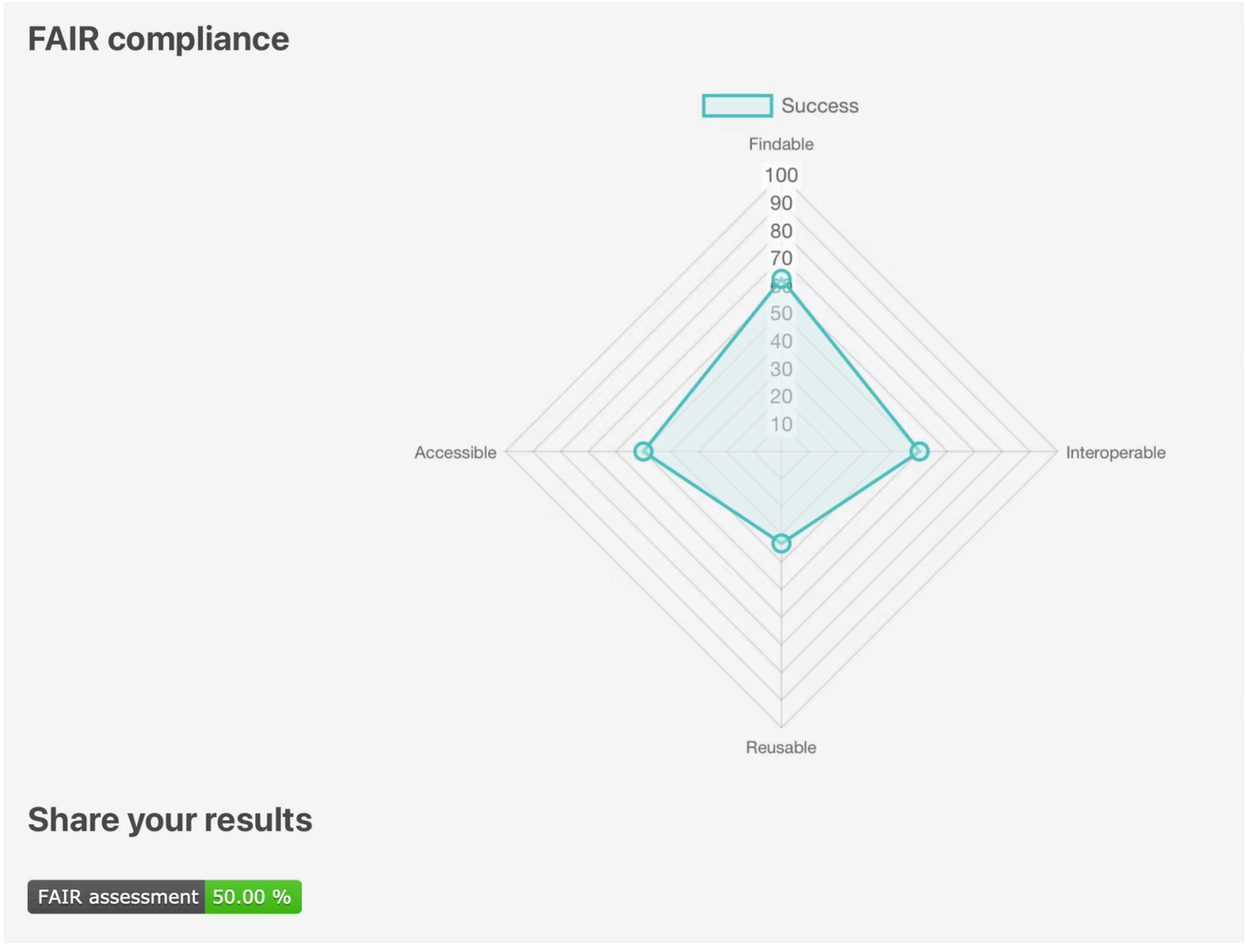
inputs & outputs ?

FAIR-Checker consumes web pages



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

FAIR-Checker produces a FAIR assessment report



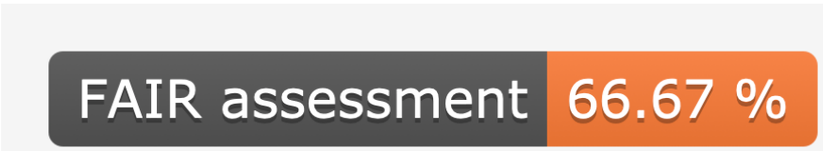
- Aggregated score per principle
- Visualised with a radar plot
- HTML badge summarising the whole evaluation

Detailed FAIR assessment results

F2B: Shared vocabularies for metadata	Check	FAIR principle F2B 2/2		
A1.1: Open resolution protocol	Check	FAIR principle A1.1 2/2		
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	
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	
I2: Use shared ontologies	Check	FAIR principle I2 2/2		
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	
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	

- 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 assessment badges



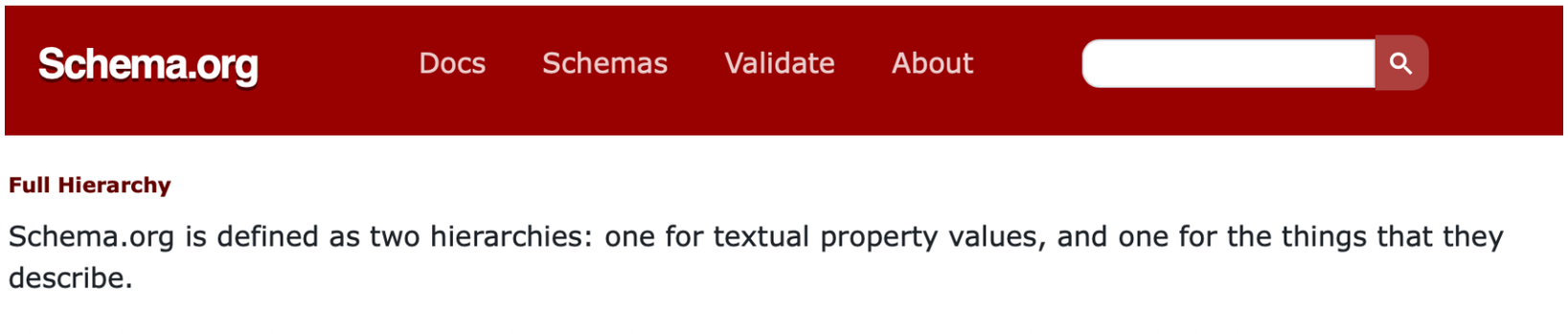

```
@prefix : <https://fair-checker.france-bioinformatique.fr/data/> .
@prefix dqv: <http://www.w3.org/ns/dqv#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

: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*)

How is collected metadata ?

1st approach, follows **search engines** recommendations



- General purpose **lightweight** ontology
- Aimed at **annotating web pages**
- Targetting **FINDABILITY**
- Originating from major search engines

Schema.org is massively adopted

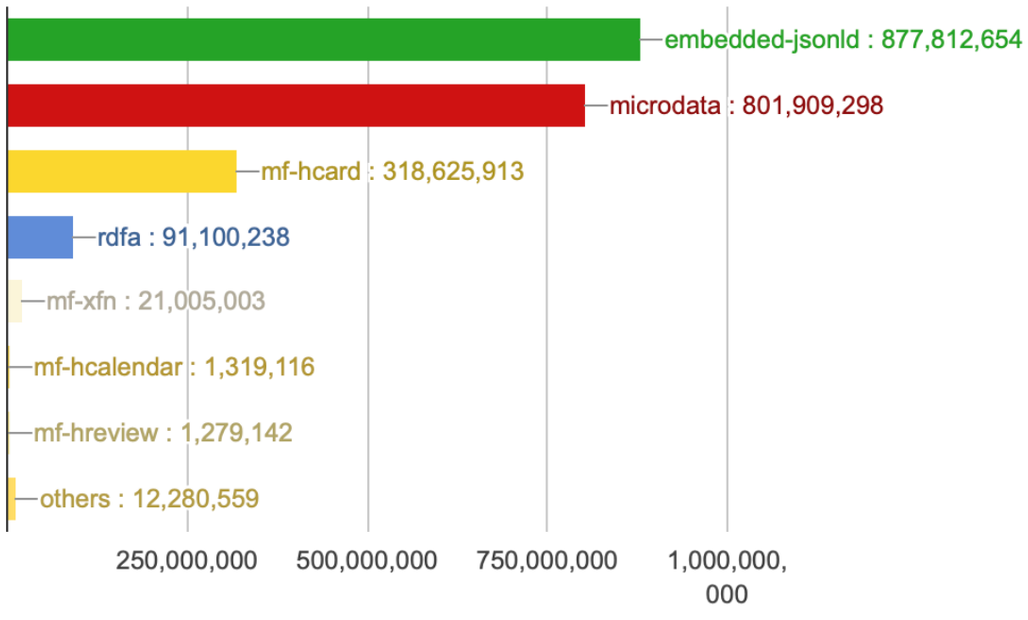
Web Data Commons

Extracting Structured Data from the Common Crawl

UNIVERSITY OF MANNHEIM

Crawl Date	October 2022	
Total Data	82.71 Terabyte	(compressed)
Parsed HTML URLs	3,048,746,652	
URLs with Triples	1,518,609,988	
Domains in Crawl	33,820,102	
Domains with Triples	14,235,035	
Typed Entities	19,072,628,514	
Triples	86,462,816,435	
Size of Extracted Data	1.6 Terabyte	(compressed)

URLs with Triples



Top Domains by Extracted Triples

1. blogspot.com (879,564,145 triples)
2. wordpress.com (458,770,038 triples)
3. wikipedia.org (190,087,065 triples)
4. yummly.com (87,112,540 triples)
5. hotels.com (81,991,039 triples)
6. boohoo.com (79,884,394 triples)
7. kayak.com (77,623,248 triples)
8. google.com (73,729,078 triples)
9. yahoo.com (65,317,838 triples)
10. southleedslife.com (63,758,451 triples)
11. indiatimes.com (58,899,559 triples)
12. freepik.com (56,124,447 triples)
13. airbnb.com (51,964,983 triples)
14. pinterest.com (47,251,484 triples)
15. soundcloud.com (45,745,317 triples)
16. apple.com (42,410,414 triples)
17. hostadvice.com (42,309,867 triples)
18. elpais.com (42,136,136 triples)
19. vsemayki.ru (38,167,517 triples)
20. smugmug.com (38,031,434 triples)
21. More

What is "understood" by search engines

Schema.org Documentation Schemas Validate About

https://workflowhub.eu/workflows/1021

Exécuter un nouveau test

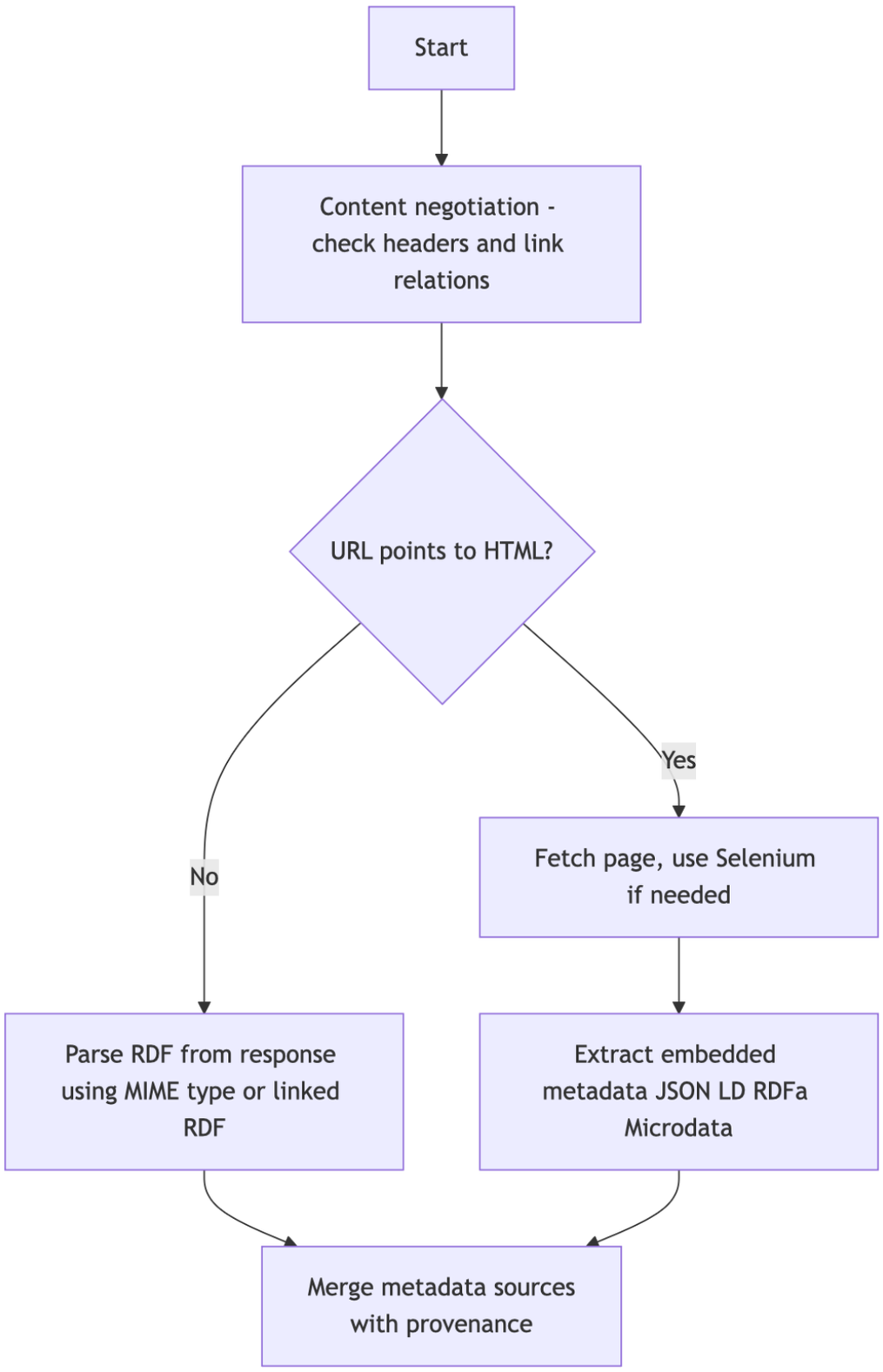

```
<meta name="csrf-param" content="authenticity_token" />
<meta name="csrf-token" content="3iMT2PTi5mzfM2t95rUKIz8Y2C8kSqwzkzc3R4sTOLCV-yIpobycXc5o6dkHjxagPyc
<script type="application/ld+json">{
"@context": "https://schema.org",
"@type": [
  "SoftwareSourceCode",
  "ComputationalWorkflow"
],
"dct:conformsTo": "https://bioschemas.org/profiles/ComputationalWorkflow/1.0-RELEASE/",
"@id": "https://workflowhub.eu/workflows/1021?version=17",
"description": "<h1>\n  \n    \n    <img alt=\"nf-core/scrnaseq\" src=\"docs/images/nf-core-scrnas
"name": "nf-core/scrnaseq",
"url": "https://workflowhub.eu/workflows/1021?version=17",
"keywords": "10x-genomics, 10xgenomics, alevin, bustools, Cellranger, kallisto, rna-seq, single-ce
"version": 17,
"license": "https://spdx.org/licenses/MIT",
"creator": [
  {
    "@type": "Person",
    "@id": "#Peter%20J%20Bailey",
    "name": "Peter J Bailey"
  },
  {
    "@type": "Person",
    "@id": "#Bailey%20PJ",
    "name": "Bailey PJ"
  },
  {
    "@type": "Person",
    "@id": "#Alexander%20Peltzer",
    "name": "Alexander Peltzer"
  },
  {
    "@type": "Person",
    "@id": "#Botvinnik%20O",
    "name": "Botvinnik O"
  },
  {
    "@type": "Person",
    "@id": "#Olga%20Botvinnik",
    "name": "Olga Botvinnik"
  },
```


found in the CITATIONS.md file.

name	nf-core/scrnaseq
url	https://workflowhub.eu/workflows/1021?version=17
url	https://workflowhub.eu/workflows/1021?version=17
keywords	10x-genomics, 10xgenomics, alevin, bustools, Cellranger, kallisto, rna-seq, single-cell, star-solo
version	17
license	https://spdx.org/licenses/MIT
license	https://spdx.org/licenses/MIT
dateCreated	2025-03-26T09:57:52+00:00
dateCreated	2025-03-26T09:57:52+00:00
dateModified	2025-03-26T09:57:53+00:00
dateModified	2025-03-26T09:57:53+00:00
creator	
@type	Person
@id	https://workflowhub.eu/workflows/1021#Peter%20J%20Bailey
name	Peter J Bailey
creator	
@type	Person
@id	https://workflowhub.eu/workflows/1021#Bailey%20PJ
name	Bailey PJ
creator	
@type	Person
@id	https://workflowhub.eu/workflows/1021#Alexander%20Peltzer
name	Alexander Peltzer
creator	
@type	Person
@id	https://workflowhub.eu/workflows/1021#Botvinnik%20O
name	Botvinnik O

→ Search engines parse RDF metadata and better "understand" the content of the web page

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

TODO

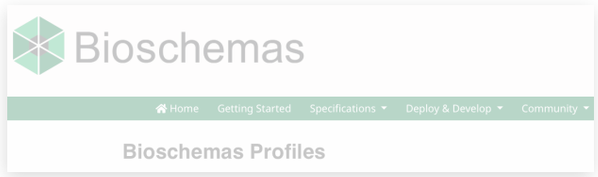
- **FAIR Signposting:** a protocol to guide machine where the metadata is stored (e.g. inside the web page, in a file on the server, at a remote location)

How biodiversity-specific rules can be defined ?

Biodiversity-specific rules ?

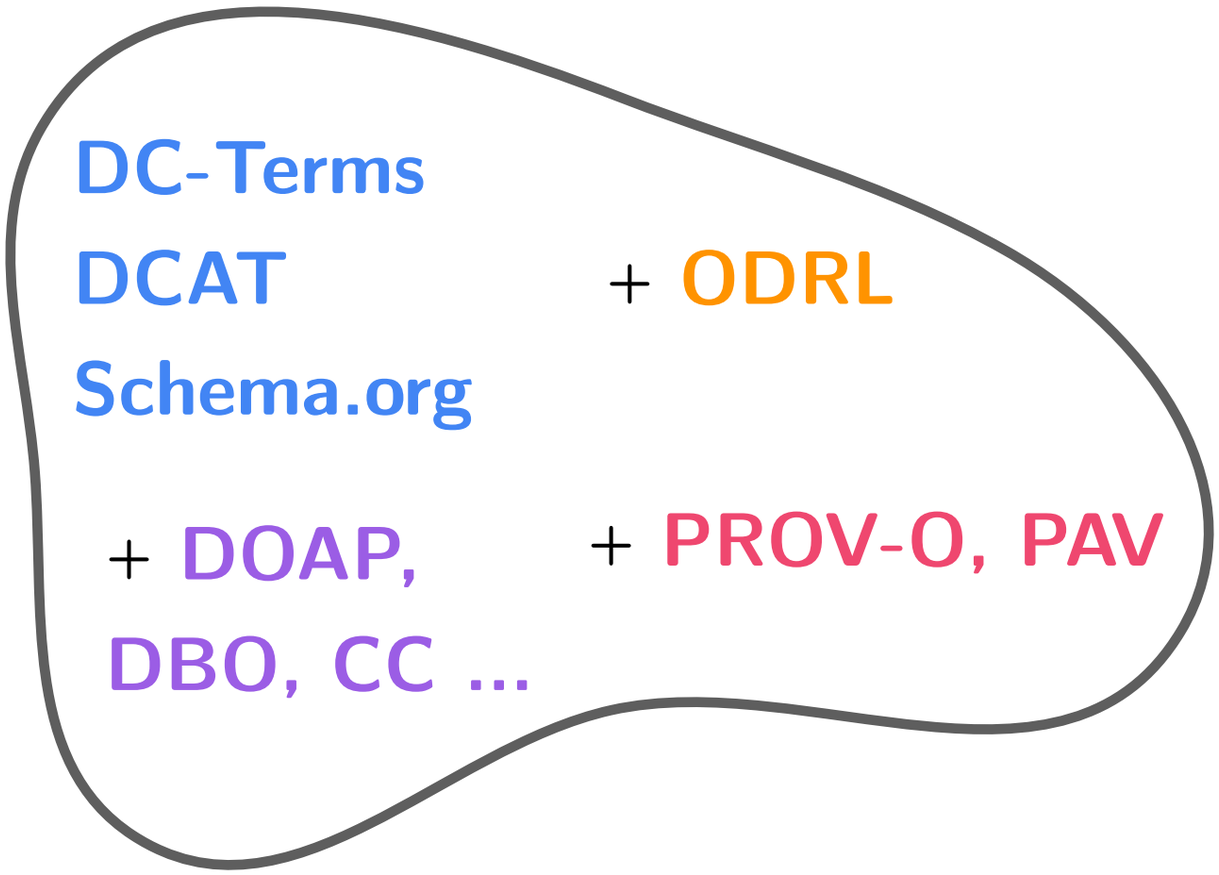
🤔 Are generic FAIR metrics enough for basic FAIR assessment of Biodiversity resources ? do we need the Biodiversity community to refine the **interpretation** and **scoring** of each metric ?

🤔 Can we reuse or extend Bioschemas profiles to increase the **quality** and **completeness** of Biodiversity metadata ?



What is evaluated to assess FAIR principles?

What is evaluated to assess FAIR principles?



What is evaluated to assess FAIR principles?

"anyOf"

Findability F1B, F2	Accessibility A1.2	Reuse (licenses) R1.1	Reuse (provenance) R1.2
dct:identifier	odrl:hasPolicy	schema:license	prov:wasGeneratedBy
schema:identifier	dct:rights	dct:license	prov:wasDerivedFrom
dct:title	dct:accessRights	doap:license	prov:wasAttributedTo
dct:description		dbo:license	prov:used
dcat:accessURL		cc:license	prov:wasInformedBy
dcat:downloadURL		xhv:license	prov:wasAssociatedWith
dcat:endpointDescription		sto:license	prov:startedAtTime
dcat:endpointURL		nie:license	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

DC-Terms

DCAT + **ODRL**

Schema.org

+ **DOAP,** + **PROV-O, PAV**

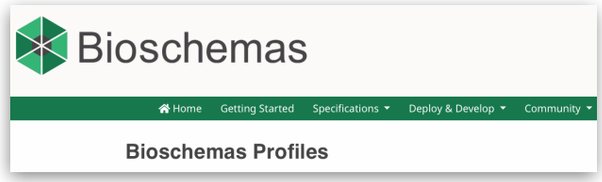
DBO, CC ...

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

Biodiversity-specific rules ?

🤔 Are generic FAIR metrics enough for basic FAIR assessment of Biodiversity resources ? do we need the Biodiversity community to refine the **interpretation** and **scoring** of each metric ?

🤔 Can we reuse or extend Bioschemas profiles to increase the **quality** and **completeness** of Biodiversity metadata ?



37 ± Life Science profiles

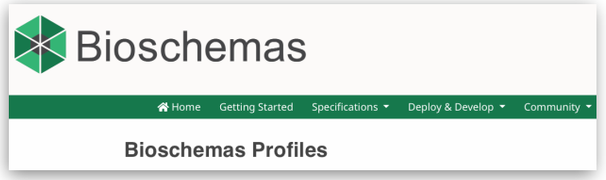
Bioschemas

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

schema.org

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

Metadata completeness



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

Marginality: Recommended.

applicationCategory	Text URL	**Schema:** Type of software application, e.g. 'Game, Multimedia'. **Bioschemas:** Type of tool e.g. Command-line tool, Web application etc. **Note:** Bioschemas have changed URL to Text in the Expected Types. This will be reverted once Bio.Tools provides stable URIs for tool types.	MANY	Please use terms from the 'Tool type' table in the biotools documentation
applicationSubCategory	Text URL	**Schema:** Subcategory of the application, e.g. 'Arcade Game'. **Bioschemas:** Use an EDAM:Topic to describe the category of application	MANY	EDAM:Topic
author	Organization Person	**Schema:** The author of this content or rating. Please note that author is special in that HTML 5 provides a special mechanism for indicating authorship via the rel tag. That is equivalent to this and may be used interchangeably.	MANY	
citation	CreativeWork URL	**Schema:** A citation or reference to another creative work, such as another publication	MANY	

↑ Back to top

```
ex:myTool   rdf:type   schema:SoftwareApplication, prov:SoftwareAgent ;
            schema:description "This tool does … " ;
            schema:license <https://spdx.org/licenses/MIT.html> ;
            schema:codeRepository <http://github.com/...> .
```

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 classes on which the profile is defined

$M = \{p_1, p_2\}$
$\rightarrow$ set of mandatory properties

$R = \{p_3, p_4, p_5\}$
$\rightarrow$ set of recommended properties

Profile → graph shapes

$P = \{C, M, R\}$
→ a metadata profile composed by target classes, mandatory and recommended properties

$C = \{C_1, C_2\}$ → set of classes on which the profile is defined

$M = \{p_1, p_2\}$
→ set of mandatory properties

$R = \{p_3, p_4, p_5\}$
→ set of 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 %}
```

Profile → graph shapes

Generated SHACL constraints for validating P

$P = \{C, M, R\}$
→ a metadata profile composed by target classes, mandatory and recommended properties

$C = \{C_1, C_2\}$ → set of classes on which the profile is defined

$M = \{p_1, p_2\}$
→ set of mandatory properties

$R = \{p_3, p_4, p_5\}$
→ set of recommended properties

Generic SHACL template

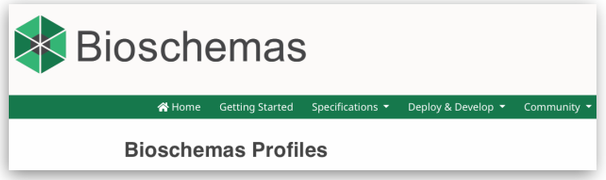
```
ns:{{shape_name}} a sh:NodeShape ;
    {% for c in target_classes %}
    sh:targetClass  {{c}} ;
    {% endfor %}

    {% 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 %}
```

→ 2 *sh:path* **strong** cardinality constraints on p_1 and p_2 **and** 3 **light** cardinality constraints on p_3, p_4 and p_5 for C_1 or C_2 instances.

Metadata completeness

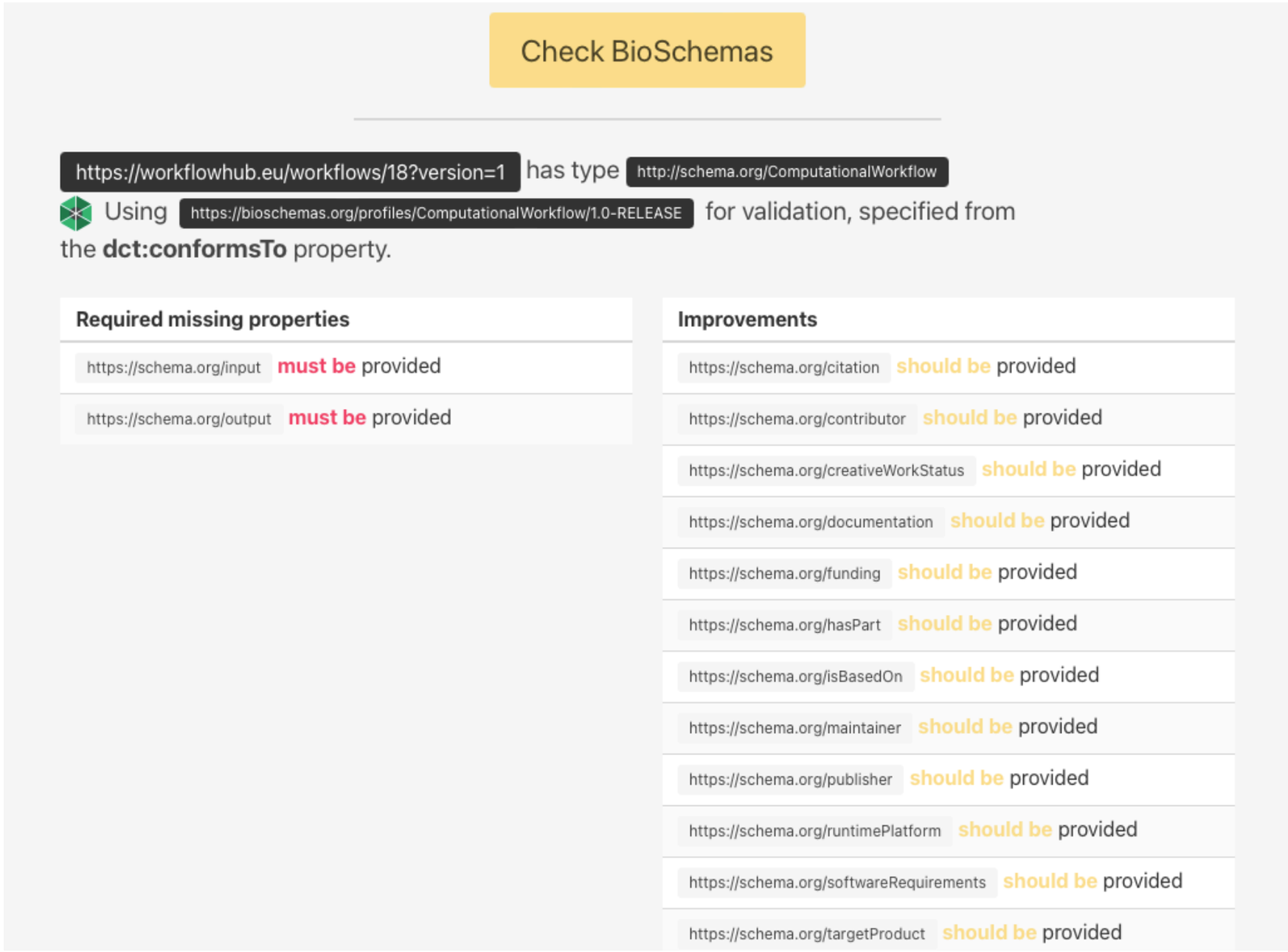


R1.3: (Meta)data meet domain-relevant community standards

Validation of Bioschemas profiles:

→ **rank missing** metadata

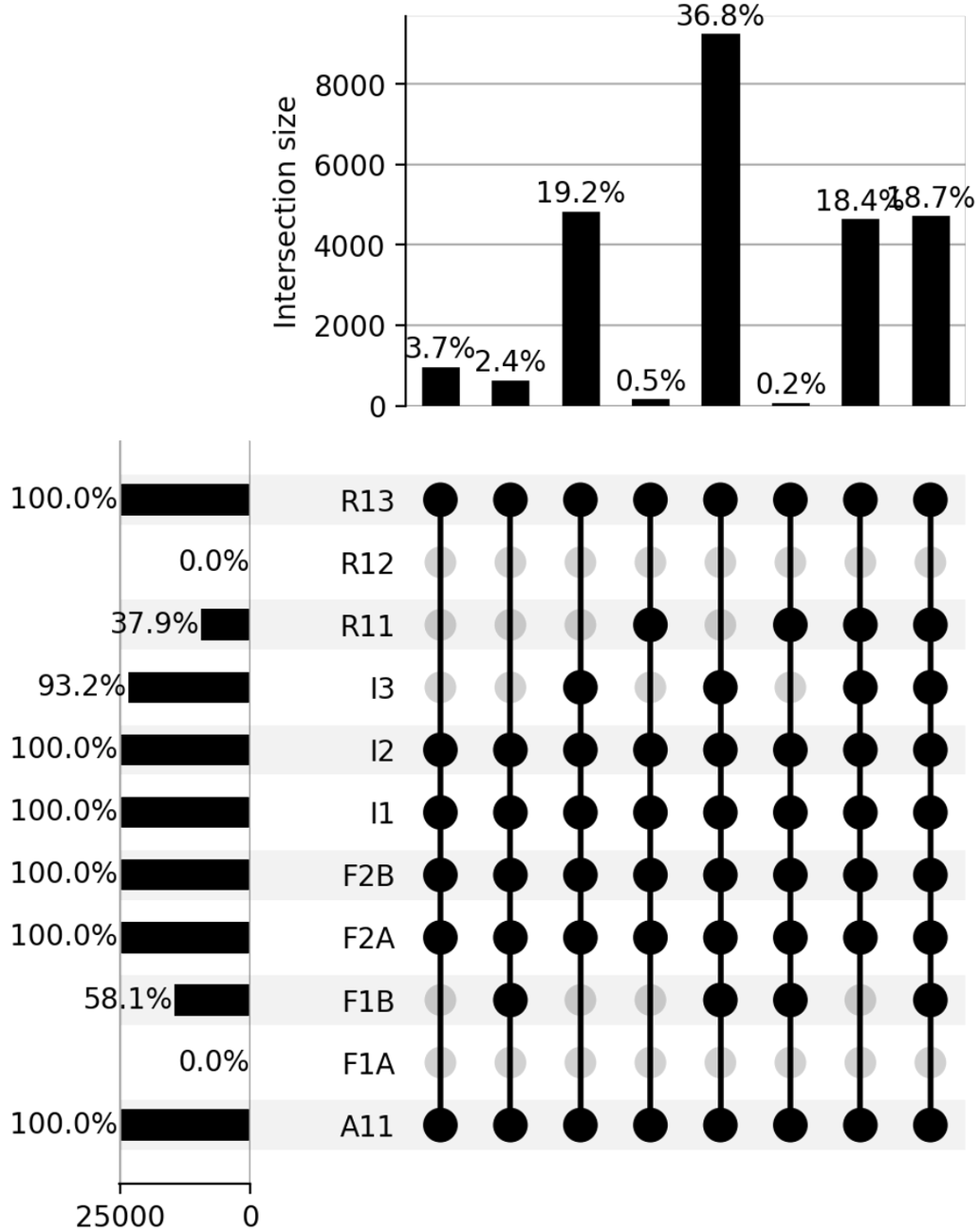
→ developer **focus** on **minimal** metadata first



What are the main **[benefits]** and limits/risks of running checks at scale?

Large-scale FAIR metrics evaluations

How FAIR are Bio.Tools registered softwares ?



Running FAIR-Checker over
25.000+ bioinformatics softwares
from Bio.tools.

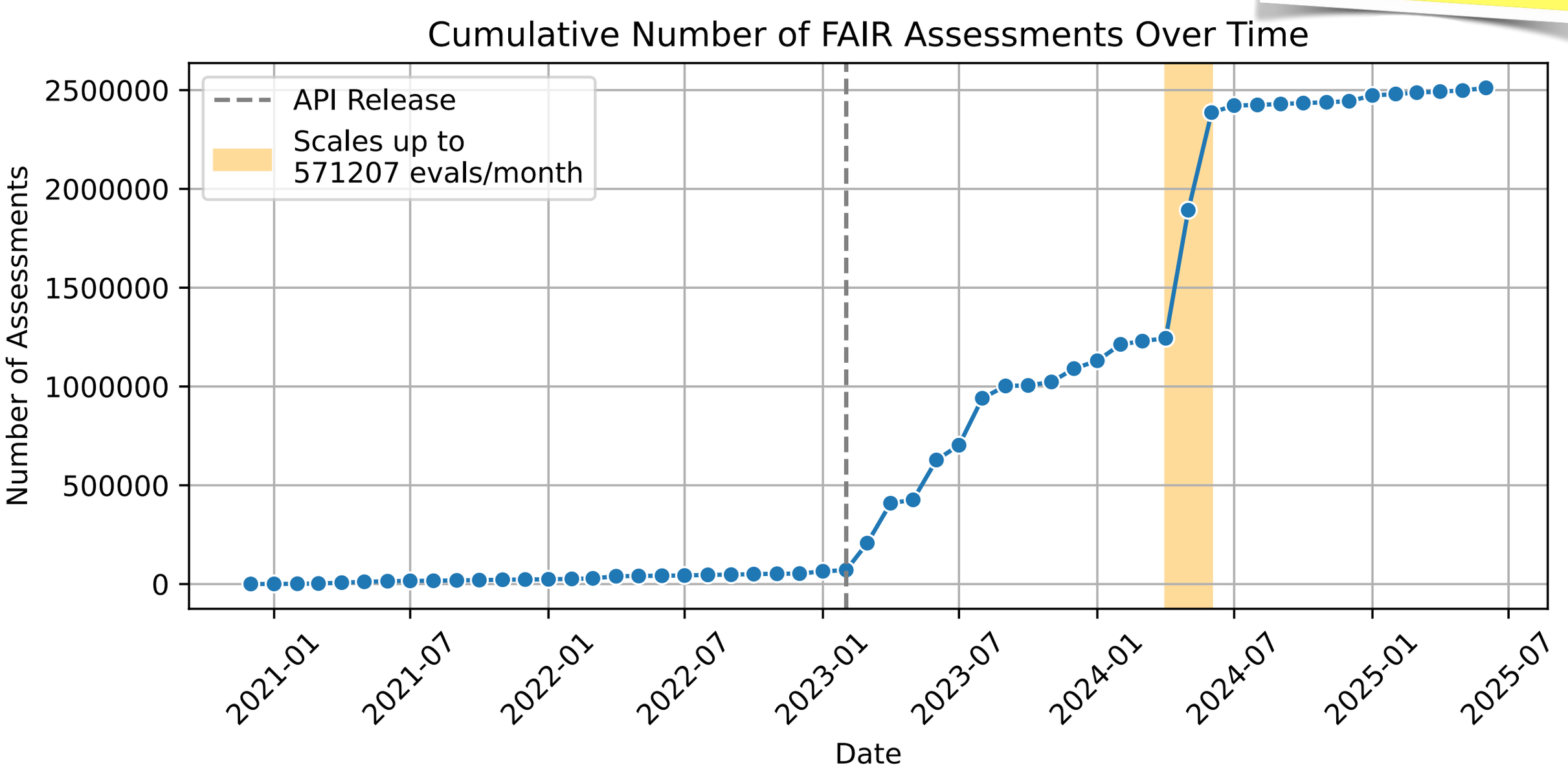
R1.1: Only 37,9% of the tools expose a **licence**

R1.2: **No provenance** metadata
→ massive impact if bio.tools developers provide PROV / PAV ontology terms

https://bio.tools/

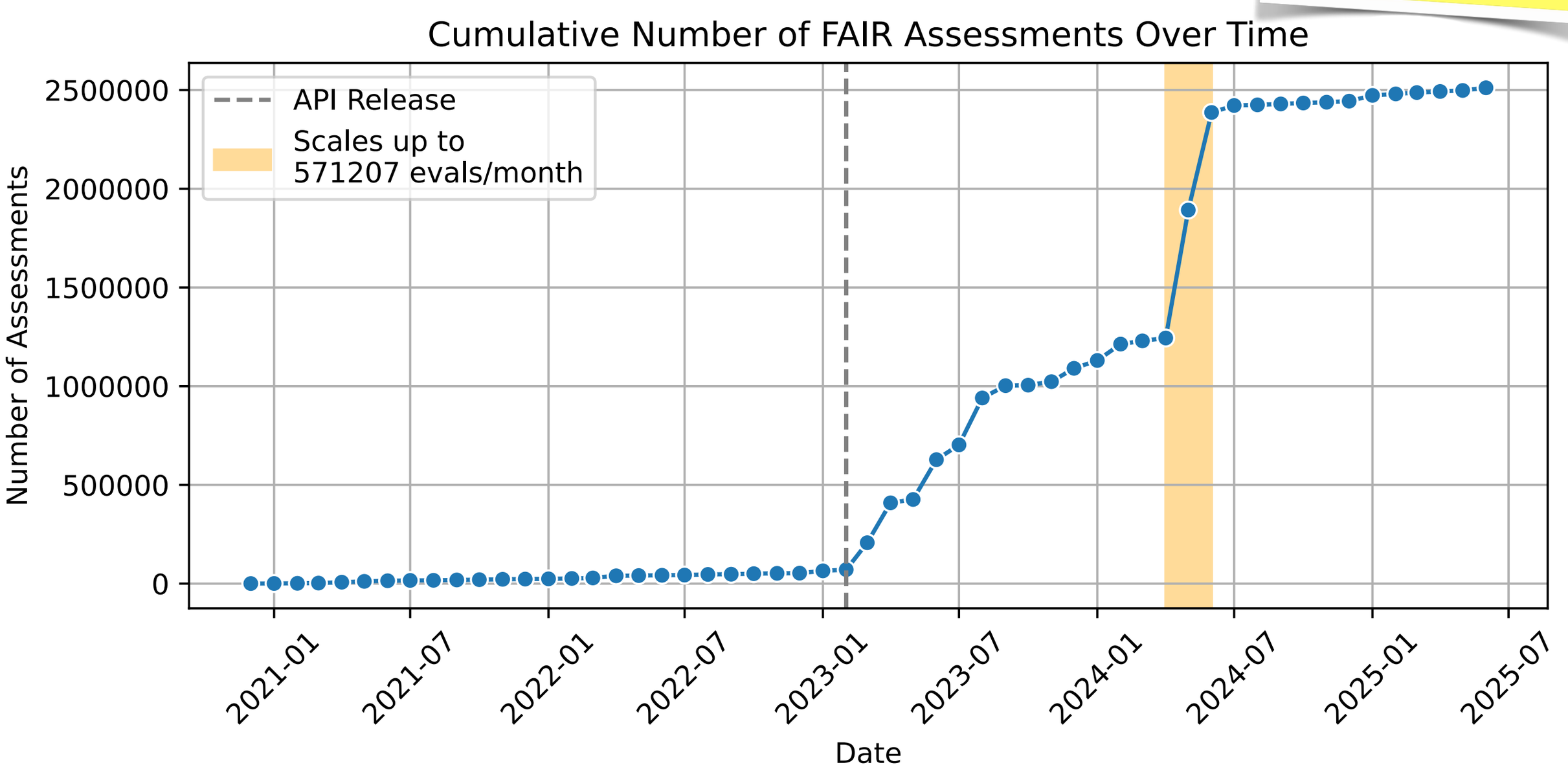
Usage statistics

not yet public,
work in progress



Usage statistics

not yet public,
work in progress



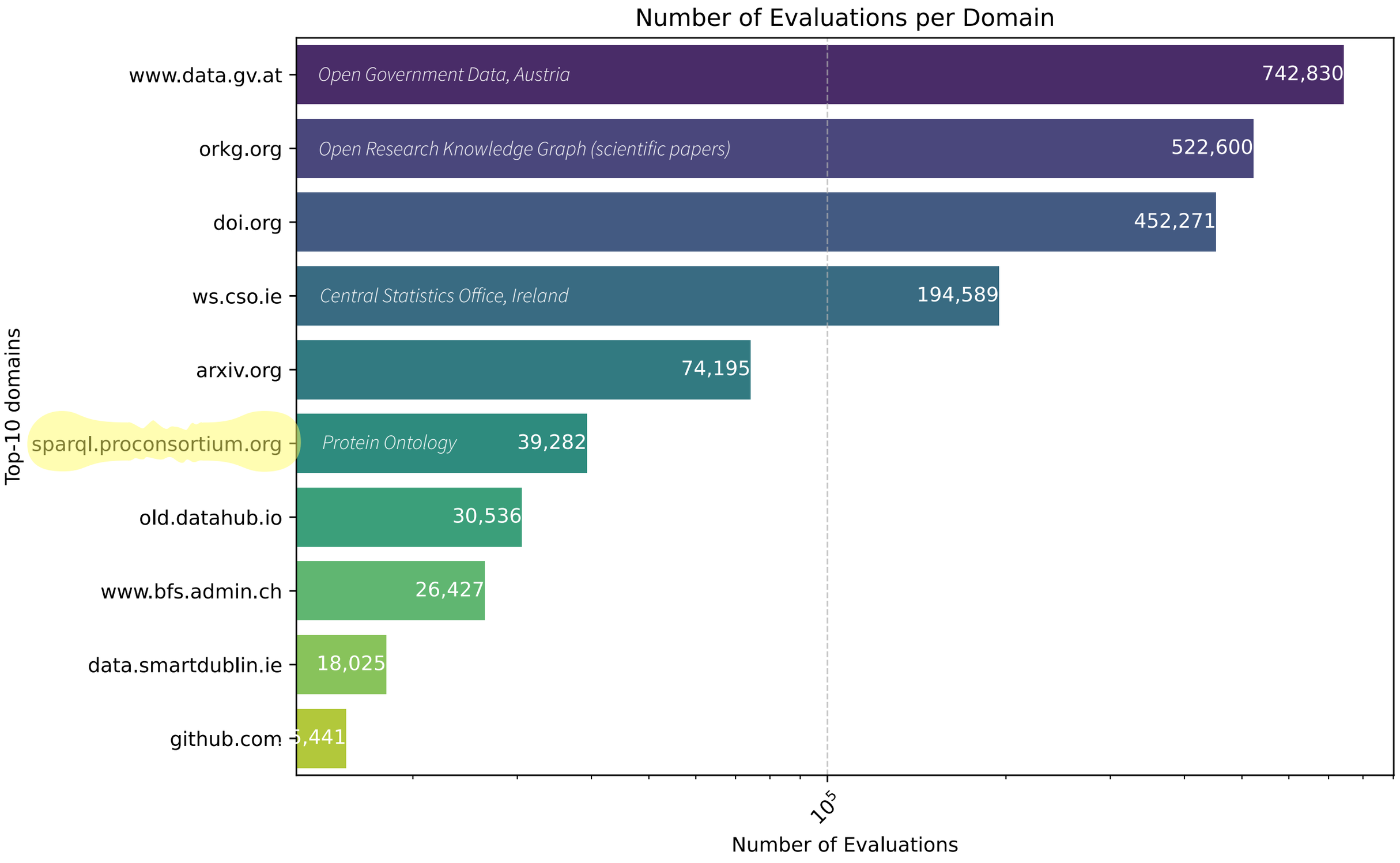
102,180 unique target URLs

47,537 unique target URLs evaluated
more than 2 times

437 unique target URLs evaluated
more than 10 times

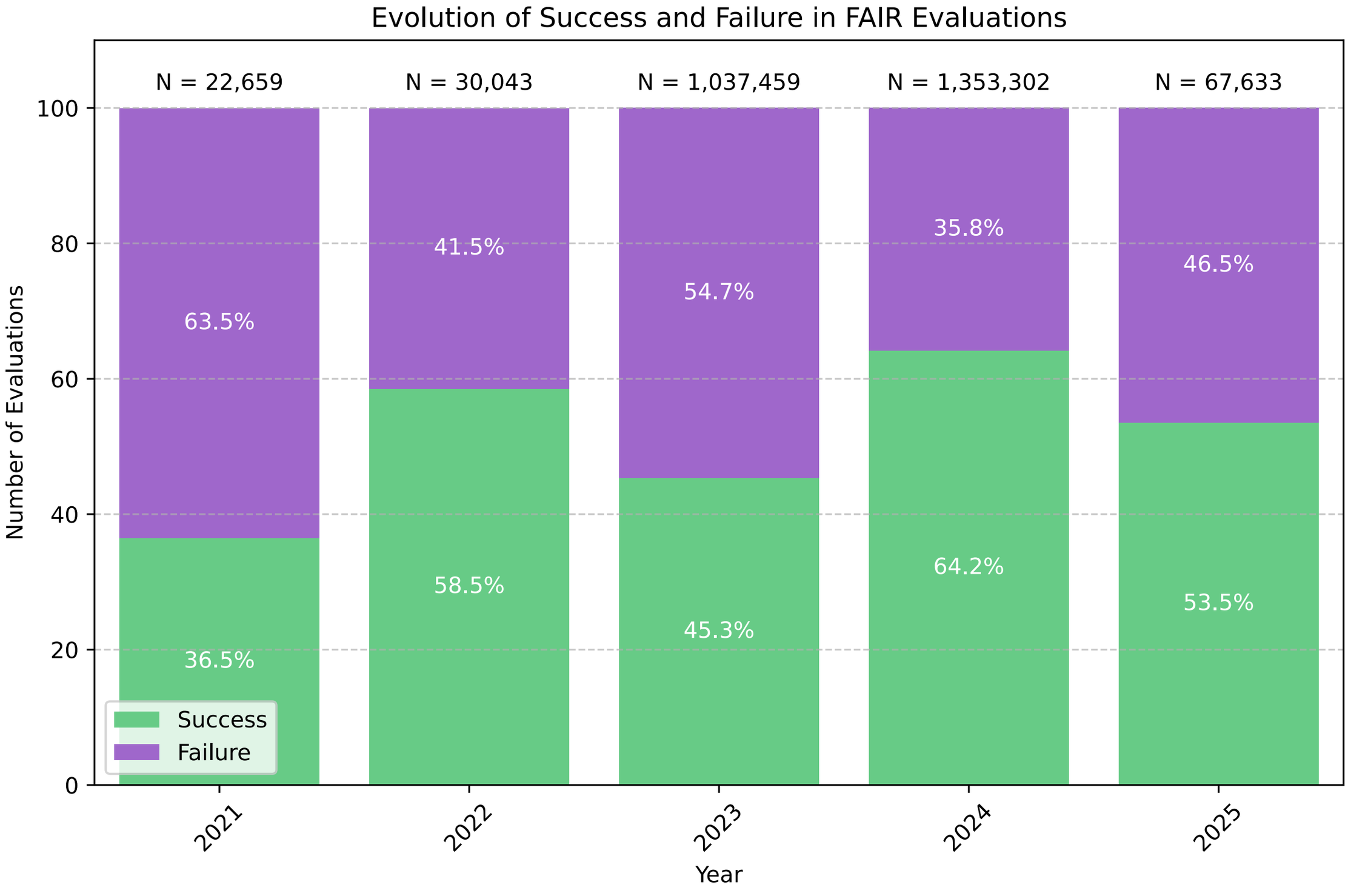
Most evaluated domains

not yet public,
work in progress



Evolution of FAIR scores

not yet public,
work in progress



Large scale evaluation risks

Large scale evaluation risks

- FAIR-Checker (FC) relies on external services for metadata quality evaluation: Bioportal, Ontology Lookup Service, Linked Open Vocabulary

Large scale evaluation risks

- FAIR-Checker (FC) relies on external services for metadata quality evaluation: Bioportal, Ontology Lookup Service, Linked Open Vocabulary
- How to avoid hammering external APIs ?
  - ⚠️ Rate limit → makes large-scale evaluations slonger
  - ✅ Caching strategy → need to "hardcode" the frequency of updates, increases FC memory consumption but small volumes
  - 🚀 Caching strategy ⊕ rate limit → faster evaluations on FC side and not

Large scale evaluation risks

- FAIR-Checker (FC) relies on external services for metadata quality evaluation: Bioportal, Ontology Lookup Service, Linked Open Vocabulary
- How to avoid hammering external APIs ?
  - ⚠️ Rate limit → makes large-scale evaluations slonger
  - ✅ Caching strategy → need to "hardcode" the frequency of updates, increases FC memory consumption but small volumes
  - 🚀 Caching strategy ⊕ rate limit → faster evaluations on FC side and not
- ❓ Do we need external services for Biodiversity specific metrics ?

Large scale evaluation risks

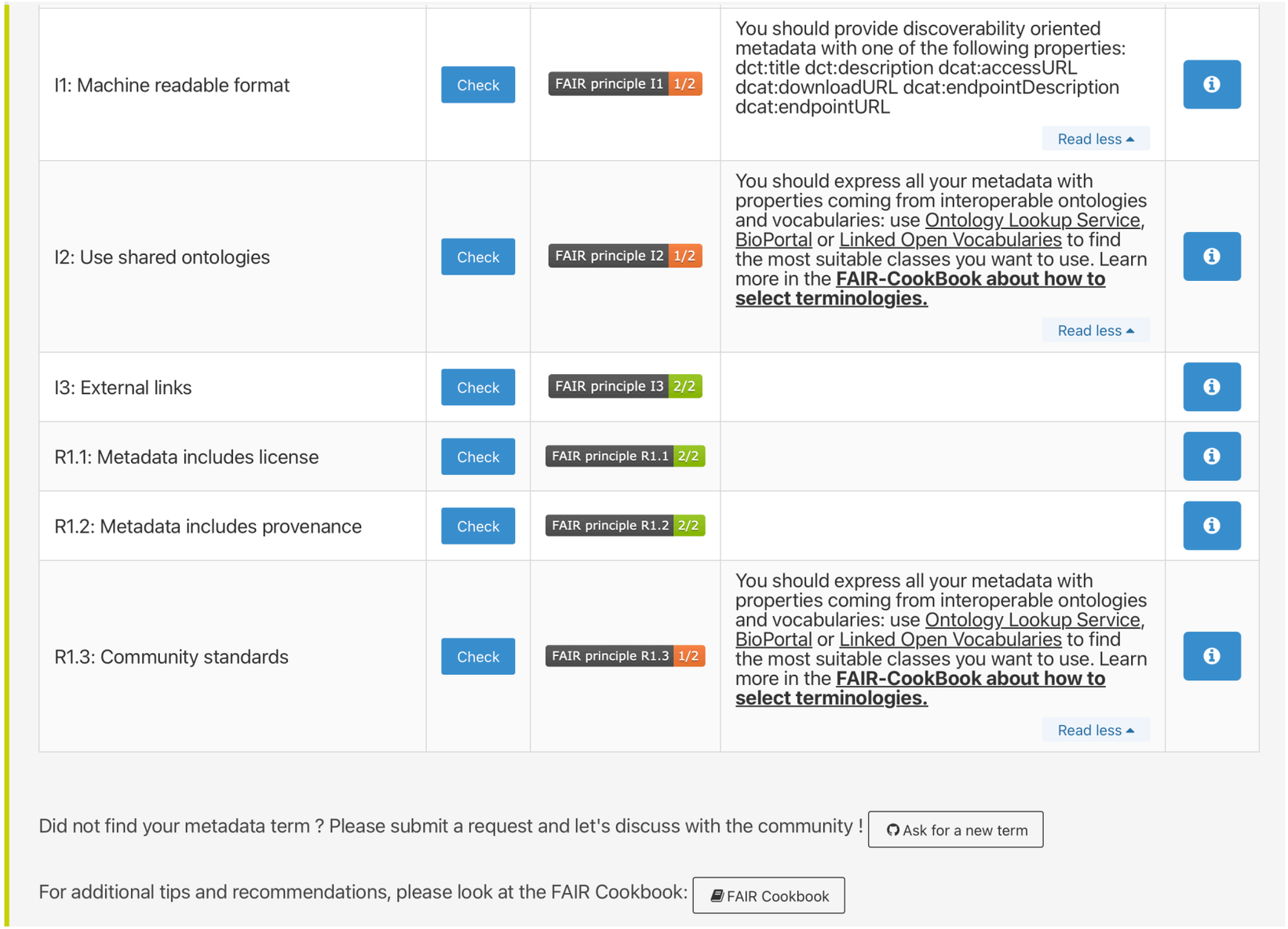
- FAIR-Checker (FC) relies on external services for metadata quality evaluation: Bioportal, Ontology Lookup Service, Linked Open Vocabulary
- How to avoid hammering external APIs ?
  - ⚠️ Rate limit → makes large-scale evaluations slonger
  - ✅ Caching strategy → need to "hardcode" the frequency of updates, increases FC memory consumption but small volumes
  - 🚀 Caching strategy ⊕ rate limit → faster evaluations on FC side and not
- ❓ Do we need external services for Biodiversity specific metrics ?
- ❓ Do we need to evaluate sensitive metadata ?
  - how to manage external authentication ? 😱
  - or provide encryption for registries working with sensitive metadata ? 👍
    - what about storing assemsent results for sensitive metadata ?
      not sure we want to address FAIRness evaluation of sensitive metadata ...

How will recommendations be generated and kept consistent?

Recommendations ≥ Score

- How a failed check becomes a **human-friendly recommendation**?
- Is it a **static text** linked to a rule?
- Does it point to **standards**, **examples**, or **documentation**?
- Can recommendations be **versioned, translated, and tailored per community**? Are these recommendations going to be stored somewhere and/or **made public**?
- This is **essential for adoption**, because communities will judge the tool by the quality of advice, not only the score.

Recommendations ≥ Score



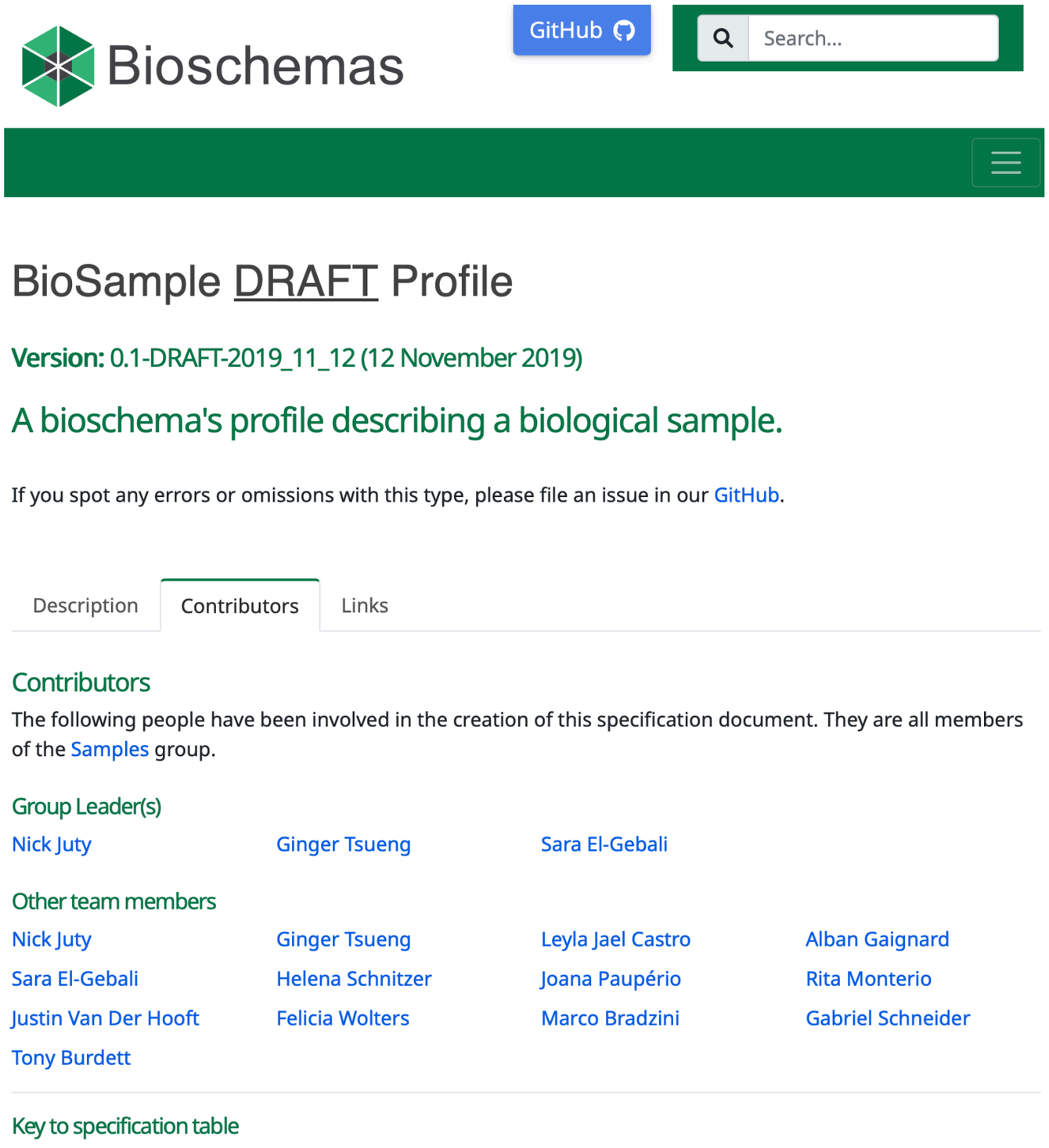
I1: Machine readable format	Check	FAIR principle I1 1/2	You should provide discoverability oriented metadata with one of the following properties: dct:title dct:description dcat:accessURL dcat:downloadURL dcat:endpointDescription dcat:endpointURL Read less ▲	ℹ
I2: Use shared ontologies	Check	FAIR principle I2 1/2	You should express all your metadata with properties coming from interoperable ontologies and vocabularies: use Ontology Lookup Service, BioPortal or Linked Open Vocabularies to find the most suitable classes you want to use. Learn more in the **FAIR-CookBook about how to select terminologies.** Read less ▲	ℹ
I3: External links	Check	FAIR principle I3 2/2		ℹ
R1.1: Metadata includes license	Check	FAIR principle R1.1 2/2		ℹ
R1.2: Metadata includes provenance	Check	FAIR principle R1.2 2/2		ℹ
R1.3: Community standards	Check	FAIR principle R1.3 1/2	You should express all your metadata with properties coming from interoperable ontologies and vocabularies: use Ontology Lookup Service, BioPortal or Linked Open Vocabularies to find the most suitable classes you want to use. Learn more in the **FAIR-CookBook about how to select terminologies.** Read less ▲	ℹ

Did not find your metadata term ? Please submit a request and let's discuss with the community ! Ask for a new term

For additional tips and recommendations, please look at the FAIR Cookbook: FAIR Cookbook

- Recommendations are in stored in a configuration file, they could be **redefined in a community-specific plugin**
- Recommendation are short and technical, with links to the **FAIR Cookbook**
- New metadata terms can be asked with **GitHub issues**

Recommendations ≥ Score



- more complex & community-oriented metadata profiles can be **authored** and **maintained** under the **Bioschemas** umbrella
- benefit from the Bioschemas **gouvernance + web site (doc, training, examples, etc.)**
- automatically translated into SHACL shapes for **validation in FAIR-Checker**

Recommendations ≥ Score

not yet public, work in progress

Let's consider these two datasets,"https://www.gbif.org/dataset/4fa7b334-ce0d-4e88-aaae-2e0c138d049e" and " https://doi.org/10.57745/VNYZ9A" which one is the most FAIR ?

AI

MCP

FAIR Checker

✓ Planning FAIR assessment using available tools

✓ Ran `perform_global_fair_assessment` – fair-checker-mcp (MCP Server)

✓ Ran `perform_global_fair_assessment` – fair-checker-mcp (MCP Server)

✓ Summarizing FAIR comparison

FAIR comparison

- The DOI landing page (https://doi.org/10.57745/VNYZ9A) is markedly more FAIR: resolvable PID, persistent identifier metadata, RDF present (163 triples), shared vocabularies detected, license and provenance present, multiple external links, and explicit access rights. Some discoverability fields (dct:title/description/accessURL) are missing, and a few properties (ogp article terms) are outside common vocabularies.
- The GBIF dataset page (https://www.gbif.org/dataset/4fa7b334-ce0d-4e88-aaae-2e0c138d049e) failed reachability for the checker (HTTP 403), so all FAIR metrics scored 0. No embedded structured metadata was detected in the crawl.

Which is most FAIR?

- The DOI record (https://doi.org/10.57745/VNYZ9A) is the more FAIR of the two.

If you want to improve the GBIF record

1. Ensure the page is fetchable without blocking automated requests (return 200).
2. Embed machine-readable metadata (JSON-LD/RDFa) with `dct:title`, `dct:description`, `dcat:downloadURL`/`accessURL`.
3. Add explicit license (`dct:license`/`schema:license`) and provenance (`dct:creator`, `prov:wasGeneratedBy`).
4. Use persistent IDs in metadata (`dct:identifier` or `schema:identifier` with the DOI/UUID).
5. Link out to at least three external domains in the metadata to improve interoperability (I3).

GPT-5.1-Codex-Max • 0.9x

Recommendations ≥ Score

not yet public, work in progress

I'm a bioinformatics researcher, considering both this tool, "http://bio.tools/jaspar" and this workflow "https://workflowhub.eu/workflows/2054" which resource is the most FAIR ? can you conclude on the FAIRness of the two hosting registries ?

AI

MCP

FAIR Checker

Comparison — key findings

- Both resources are reachable and have persistent identifiers (F1A=2, F1B=2).
- Both expose structured, machine-readable metadata (RDF triples): bio.tools JASPAR ~231 triples, WorkflowHub workflow ~41 triples.
- Metadata completeness: both lack some discoverability fields (dct:title/dct:description/accessURL) per the checker (F2A and I1 weak=1).
- Vocabularies & interoperability:
  - bio.tools: most RDF classes/properties map to OLS/LOV/BioPortal; a few bio.tools-specific or OGP properties are outside registries (I2 mostly OK).
  - WorkflowHub: most properties map to registries, but `schema:ComputationalWorkflow` was flagged as not known by OLS/LOV (I2 warning).
- External links, license, provenance, access rights:
  - bio.tools: many external domains (I3=2), license and provenance present (R1.1/R1.2=2), access policy present (A1.2=2).
  - WorkflowHub: multiple external domains (I3=2), license and provenance present (R1.1/R1.2=2), access policy present (A1.2=2).
- Community standards (R1.3): both score weakly (1) — RDF uses known vocabularies but contains some properties/classes outside OLS/LOV.

Conclusion — which is more FAIR?

- Overall both are similar and reasonably FAIR for machine-actionability and reusability. bio.tools has a stronger RDF footprint (more triples) and broader external linking; WorkflowHub has fewer triples but equivalent scores on license/provenance/access. Neither is perfect — both miss some discoverability fields and use a few non-registered properties.
- If forced to pick a winner: bio.tools/jaspar is marginally more FAIR due to richer RDF metadata and broader external linking.

Conclusions about the registries

- bio.tools: generally FAIR-aware — publishes rich RDF metadata, uses persistent identifiers, includes license/provenance, and links widely. Minor issues: some site-specific properties and Open Graph properties are outside ontology registries; add standard discoverability fields (dct:title/description/accessURL) to improve F2A/I1.
- WorkflowHub: FAIR-minded registry with structured metadata, persistent IDs, license/provenance and external links. It uses community vocabularies mostly but has a small mismatch (`schema:ComputationalWorkflow` not found in OLS/LOV) and slightly less RDF richness than bio.tools. Adding or aligning classes with registries (or registering them) and improving discoverability fields would help.

Recommendations ≥ Score

not yet public, work in progress

I'm a bioinformatics researcher, considering both this tool, "http://bio.tools/jaspar" and this workflow "https://workflowhub.eu/workflows/2054" which resource is the most FAIR ? can you conclude on the FAIRness of the two hosting registries ?

AI

MCP

FAIR Checker

Comparison — key findings

- Both resources are reachable and have persistent identifiers (F1A=2, F1B=2).
- Both expose structured, machine-readable metadata (RDF triples): bio.tools JASPAR ~231 triples, WorkflowHub workflow ~41 triples.
- Metadata completeness: both lack some discoverability fields (dct:title/dct:description/accessURL) per the checker (F2A and I1 weak=1).
- Vocabularies & interoperability:
  - bio.tools: most RDF classes/properties map to OLS/LOV/BioPortal; a few bio.tools-specific or OGP properties are outside registries (I2 mostly OK).
  - WorkflowHub: most properties map to registries, but `schema:ComputationalWorkflow` was flagged as not known by OLS/LOV (I2 warning).
- External links, license, provenance, access rights:
  - bio.tools: many external domains (I3=2), license and provenance present (R1.1/R1.2=2), access policy present (A1.2=2).
  - WorkflowHub: multiple external domains (I3=2), license and provenance present (R1.1/R1.2=2), access policy present (A1.2=2).
- Community standards (R1.3): both score weakly (1) — RDF uses known vocabularies but contains some properties/classes outside OLS/LOV.

Conclusion — which is more FAIR?

- Overall both are similar and reasonably FAIR for machine-actionability and reusability. bio.tools has a stronger RDF footprint (more triples) an
- If f to

Conclu

- bio pe Mi pr dis F2
- Wo pe co (s

less RDF richness than bio.tools. Adding or aligning classes with registries (or registering them) and improving discoverability fields would help.

The following classes are not found in registries:

- http://schema.org/ComputationalWorkflow

Some terms have not been found in LOV, OLS, nor BioPortal.
You can use these semantics registries to search for already existing synonyms terms, or to propose new terms (Classes or Properties) to the community

Classes	Properties
http://schema.org/ComputationalWorkflow OLS LOV BioPortal	http://purl.org/dc/terms/conformsTo OLS LOV BioPortal
http://schema.org/ComputerLanguage OLS LOV BioPortal	http://schema.org/creator OLS LOV BioPortal
http://schema.org/Organization OLS LOV BioPortal	http://schema.org/dateCreated OLS LOV BioPortal
http://schema.org/Person OLS LOV BioPortal	http://schema.org/dateModified OLS LOV BioPortal
http://schema.org/Project OLS LOV BioPortal	http://schema.org/description OLS LOV BioPortal
http://schema.org/SoftwareSourceCode OLS LOV BioPortal	http://schema.org/identifier OLS LOV BioPortal

Milestones and deliverables

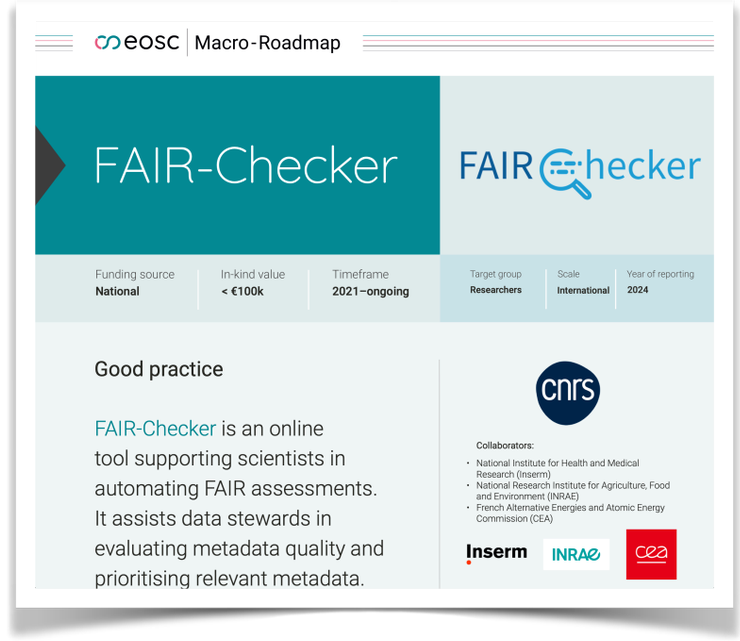
Milestones and deliverable for the FAIR-Checker dev task

Task T2. Implement biodiversity-specific FAIR assessment metrics through a FAIR-Checker biodiversity plugin

Milestone & Deliverables	Type	Contributors	Date
M2.1 A FAIRness evaluation of representative entries in ENA, BioSample, and DiSSCo	Report	Elixir-FR, Elixir-CH, Elixir-UK	T0+2
D2.1 Biodiv FAIR metrics implementation	Software	Elixir-FR,Elixir-UK	T0+7
D2.2 Biodiv metadata profile (Bioschemas profile + SHACL rules)	Specification	Elixir-UK,Elixir-FR, Elixir-CH	T0+9
D2.3 FAIR-Checker biodiversity plugin	Software	Elixir-FR, Elixir-CH, Elixir-UK	T0+12

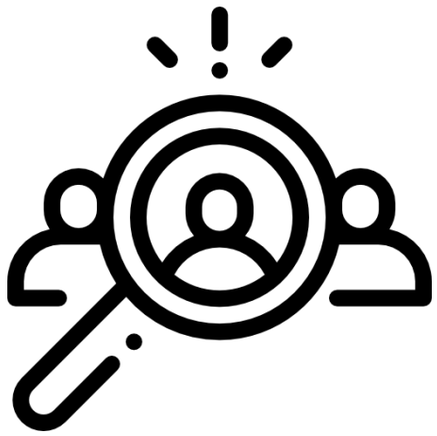
A community service, with many improvements

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



Future works

- Support "FAIR-Signposting" for **better metadata consumption**
- **Extensibility trough "plugins" (Biodiversity plugin)**
- Bioschemas **profile recommender**
- Allow users to test **missing metadata**
- Retrospective **usage study**
- **Permanent IDs** (e.g. https://w3id.org/fairchecker/data/66f682517e5dc5bcb9430aef)
- MCP server: **interaction with LLM agents**
- Suggest semantic metadata based on **AI generation** pipeline



Additional resources

- W3C RDF Primer (https://www.w3.org/TR/rdf11-primer/)
- W3C JSON-LD Primer (https://json-ld.org/primer/latest/)
- JSON-LD playground (https://json-ld.org/playground/)
- schema.org validator (https://validator.schema.org/)
- Gaignard, A., Rosnet, T., de Lamotte, F., Lefort, V., & Devignes, M. (2023). ***FAIR-Checker: supporting digital resource findability and reuse with Knowledge Graphs and Semantic Web standards***. Journal of Biomedical Semantics, 14. https://doi.org/10.1186/s13326-023-00289-5
- Lamarre, P., Andersen, J., Gaignard, A., Cazalens, S. (2025). ***A Deep Dive into FAIRness Assessment: UReFM, a Formal Framework for Representing, Analyzing and Comparing Measures***. In: Hameurlain, A., Tjoa, A.M. (eds) Transactions on Large-Scale Data- and Knowledge-Centered Systems LVIII. Lecture Notes in Computer Science(), vol 16080. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-72116-2_4

Supplementary slides

Reuse of ontologies

F4: (Meta)data are registered or indexed in a searchable resource

I2: (Meta)data use vocabularies that follow the FAIR principles

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

Step 3: Metadata quality checks

Controlled vocabularies | Bioschemas

We now have a Knowledge Graph grounded to ontology concepts (classes) and relations (properties). Are these classes and properties already known in reference ontology registries such as LOV, OLS or BioPortal ?

Check Vocabularies

Congratulations ! All Classes and Properties are referenced in one or more of the registries checked !

Classes			
http://schema.org/DataDownload	OLS	LOV	BioPortal
http://schema.org/Organization	OLS	LOV	BioPortal
http://schema.org/Person	OLS	LOV	BioPortal
https://schema.org/Dataset	OLS	LOV	BioPortal

Properties			
http://ogp.me/ns#description	OLS	LOV	BioPortal
http://ogp.me/ns#site_name	OLS	LOV	BioPortal
http://ogp.me/ns#title	OLS	LOV	BioPortal
http://ogp.me/ns#url	OLS	LOV	BioPortal
http://schema.org/affiliation	OLS	LOV	BioPortal
http://schema.org/author	OLS	LOV	BioPortal
http://schema.org/contentSize	OLS	LOV	BioPortal
http://schema.org/contentUrl	OLS	LOV	BioPortal
http://schema.org/creator	OLS	LOV	BioPortal
http://schema.org/dateCreated	OLS	LOV	BioPortal
http://schema.org/dateModified	OLS	LOV	BioPortal
http://schema.org/datePublished	OLS	LOV	BioPortal

Zenodo dataset "Comparison of spatial transcriptomics technologies used for tumor cryosections" https://zenodo.org/records/10863259

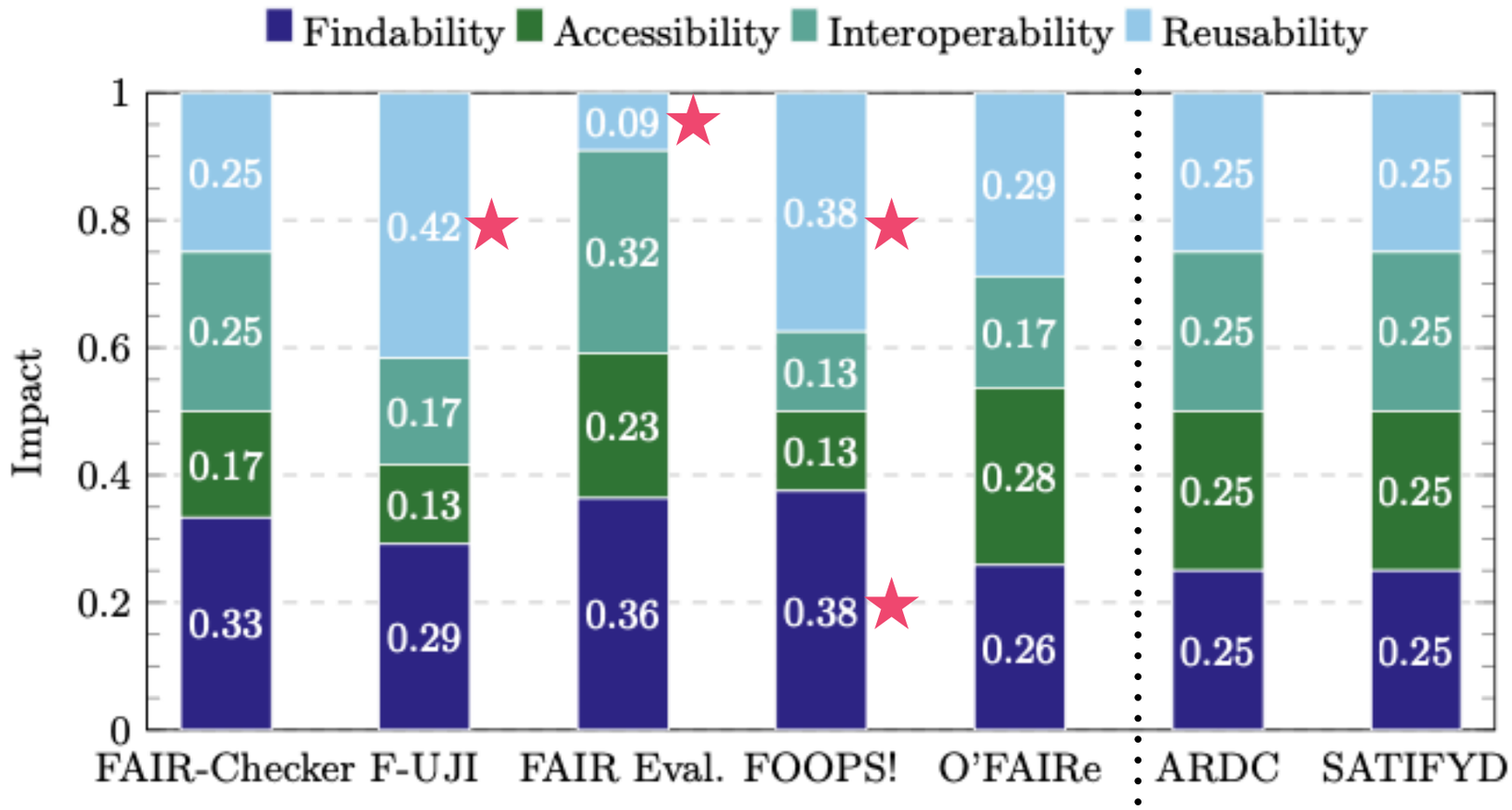
Do engines reach consensus on FAIR assessment ?

Resource	F-UJI (%)	FAIR-Checker (%)	Std dev
Dataset (PANGAEA) [31]	91	91.70	0.49
Gene Ontology (OLS) [21]	18	16.70	0.92
Dataset (Harvard Dataverse) [23]	75	79.20	2.97
Dataset (Kaggle) [26]	60	70.80	7.64
Online course (Moodle) [28]	4	16.70	8.98
Dataset (Governmental platform) [22]	52	70.80	13.29
Dataset (WHO) [39]	27	50.00	16.26
Training material (TeSS) [36]	39	70.80	22.49
Bioinformatics tool (bio.tools) [6]	18	54.20	25.60
Dataset (RDF metadata) [33]	43	87.50	31.47

- Higher scores for FAIR-Checker
- Last two entries: std. dev. > 25 % ?

How much biased are FAIR assesment tools ?

Are all principle equally contributing to the global FAIR assesment score ?



→ *How to get a good FAIR score with a minimal effort ?*

- Pay attention to identifiers (F), license + provenance + domain-specific standards (R) if you use FOOPS!
- Not useful to spend energy on provenance or domain ontologies if you use FAIR-Evaluator ...
- ... but pay attention to it if you use F-UJI.