

# (more) Findable bionformatics softwares with Bioschemas

*"machine actionable Software Management Plans" workshop,  
Cologne, 31/05/2023*

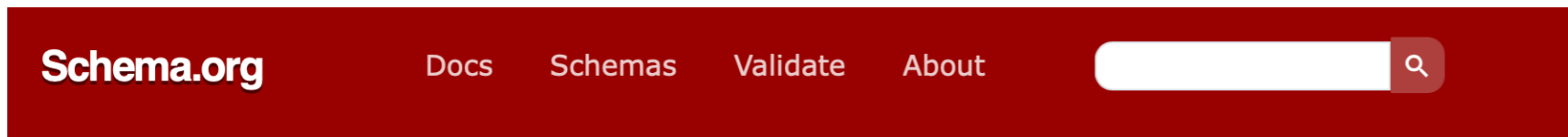
Alban Gaignard  
CNRS, ELIXIR-FR  
Institut du Thorax, Nantes, France



Bioschemas ?

[schema.org](https://schema.org) in Life Sciences

# schema.org



## Full Hierarchy

Schema.org is defined as two hierarchies: one for textual property values, and one for the things that they describe.

This is the main schema.org hierarchy: a collection of types (or "classes"), each of which has one or more parent types. Although a type may have more than one super-type, here we show each type in one branch of the tree only. There is also a parallel hierarchy for **data types**.

## Types:

[Close hierarchy](#) / [Open hierarchy](#)

- Thing
  - ▶ Action +
  - ▶ BioChemEntity +
  - ▶ CreativeWork +
  - ▶ Event +
  - ▶ Intangible +
  - ▶ MedicalEntity +
  - ▶ Organization +
  - ▶ Person +
  - ▶ Place +
  - Product
    - DietarySupplement
    - Drug
    - IndividualProduct
    - ProductCollection
    - ProductGroup

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



# Schema.org is massively adopted

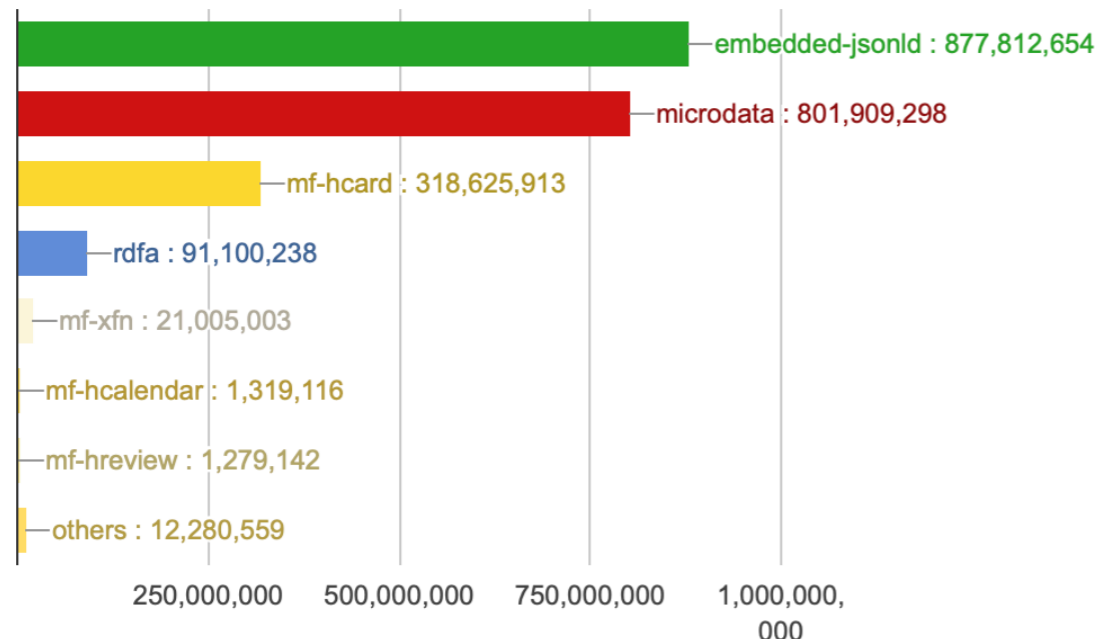
## Web Data Commons

Extracting Structured Data from the Common Crawl



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

**Schema.org for  
Life-Science resources ?**

# 37 ± Life Science profiles



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

- ▶ different use of schema.org classes and properties
- ▶ Communities agree on minimal/recommended/optional annotation

# Bioschemas profiles

*Profiles ≠ Classes (types)*

Bioschemas **profiles** specify which RDF triples are expected to describe specific entities :

- which ontology classes or properties should be used (mostly from Schema.org)
- different marginalities / priorities (minimal, recommended, optional)
- different cardinalities (one or many) for predicates

Why it's an  
important topic ?

1. Community agreement on  
metadata fields to focus on





# Computational tool profile

**Version:** 1.0-RELEASE (11 October 2021)

**Bioschemas specification for describing a SoftwareApplication in the Life Sciences**

If you spot any errors or omissions with this type, please file an issue in our [GitHub](#).

Description

Contributors

Links

## Schema.org hierarchy






This Profile fits into the schema.org hierarchy as follows:

[Thing](#) > [CreativeWork](#) > [SoftwareApplication](#)

```
http://my.tool/    rdf:type    schema:SoftwareApplication .
```

# Example

[...]

Property	Expected Type	Description	CD	Controlled Vocabulary	Example
<b>Marginality: Minimum.</b>					
<a href="#">@context</a>	<a href="#">URL</a>	Used to provide the context (namespaces) for the JSON-LD file. Not needed in other serialisations.	ONE		
<a href="#">@type</a>	<a href="#">Text</a>	Schema.org/Bioschemas class for the resource declared using JSON-LD syntax. For other serialisations please use the appropriate mechanism. While it is permissible to provide multiple types, it is preferred to use a single type.	MANY	Schema.org, Bioschemas	
<a href="#">@id</a>	<a href="#">IRI</a>	Used to distinguish the resource being described in JSON-LD. For other serialisations use the appropriate approach.	ONE		
<a href="#">dct:conformsTo</a>	<a href="#">IRI</a>	Used to state the Bioschemas profile that the markup relates to. The versioned URL of the profile must be used. Note that we use a CURIE in the table here but the full URL for Dublin Core terms must be used in the markup ( <a href="http://purl.org/dc/terms/conformsTo">http://purl.org/dc/terms/conformsTo</a> ), see example.	ONE	Bioschemas profile versioned URL	
<a href="#">description</a>	<a href="#">Text</a>	<b>Schema:</b> A description of the item.  <b>Bioschemas:</b> A short description of the tool.	ONE		

[...]

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

## 2. Semantic search



# Bioschemas + EDAM → Knowledge Graph

<https://biohackrxiv.org/79kje/>

## Query 6: Top-10 most represented EDAM operations

```
SELECT ?operation (COUNT(?operation) as ?count) ?label WHERE {  
  ?x rdf:type <http://schema.org/SoftwareApplication> ;  
  <http://schema.org/name> ?name ;  
  <http://schema.org/featureList> ?operation .  
  ?operation rdfs:label ?label .  
}  
GROUP BY ?operation ?label  
ORDER BY DESC(?count)  
LIMIT 10
```

### SPARQL | HTML5 table

operation	count	label
<a href="http://edamontology.org/operation_0337">http://edamontology.org/operation_0337</a>	1783	"Visualisation"
<a href="http://edamontology.org/operation_3435">http://edamontology.org/operation_3435</a>	1346	"Standardisation and normalisation"
<a href="http://edamontology.org/operation_3196">http://edamontology.org/operation_3196</a>	1208	"Genotyping"
<a href="http://edamontology.org/operation_2422">http://edamontology.org/operation_2422</a>	1181	"Data retrieval"
<a href="http://edamontology.org/operation_2495">http://edamontology.org/operation_2495</a>	1127	"Expression analysis"
<a href="http://edamontology.org/operation_2421">http://edamontology.org/operation_2421</a>	958	"Database search"
<a href="http://edamontology.org/operation_0224">http://edamontology.org/operation_0224</a>	956	"Query and retrieval"
<a href="http://edamontology.org/operation_3659">http://edamontology.org/operation_3659</a>	805	"Regression analysis"
<a href="http://edamontology.org/operation_3891">http://edamontology.org/operation_3891</a>	777	"Essential dynamics"
<a href="http://edamontology.org/operation_3799">http://edamontology.org/operation_3799</a>	773	"Quantification"

- ▶ **Instrumented** bio.tools **registry** to produce Bioschemas markup  
→ cost of annotating a software = cost of publishing it through bio.tools
- ▶ **Query** able knowledge graph  
→ **SPARQL** endpoint  
+ possibly part of other federated queries (e.g. OpenCitation)



# 3. Improved FAIRness of software tools



# FAIR-Checker

## Problem statement

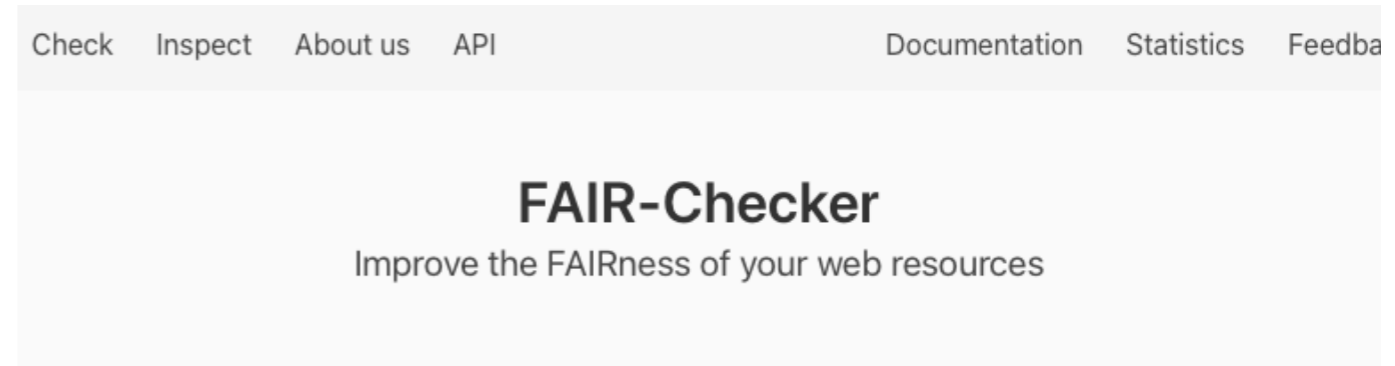
FAIR principles

≠ technical specifications

Semantic web technologies +  
knowledge graphs can **operate FAIR**  
assessment, but require specific skills

## Objectives

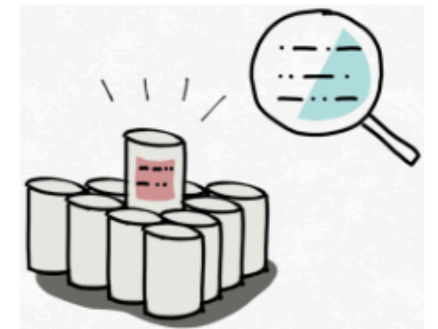
- (i) monitor the **results** of FAIR  
metrics evaluation
- (ii) improve the **quality** of  
embedded metadata



## Welcome

FAIR-Checker is a tool aimed at assessing FAIR principles and empowering data provider to enhance the quality of their digital resources.

Data providers and consumers can **check** how FAIR are web resources. Developers can explore and **inspect** metadata exposed in web resources.

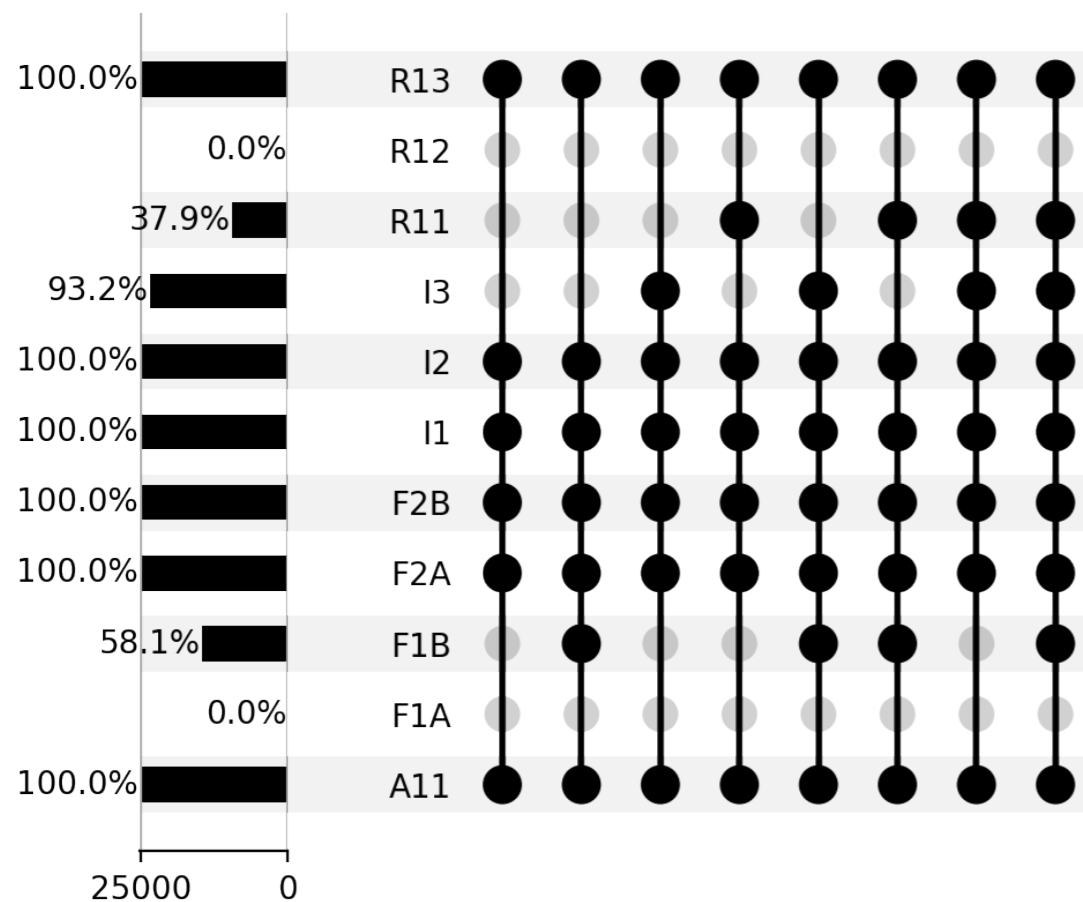
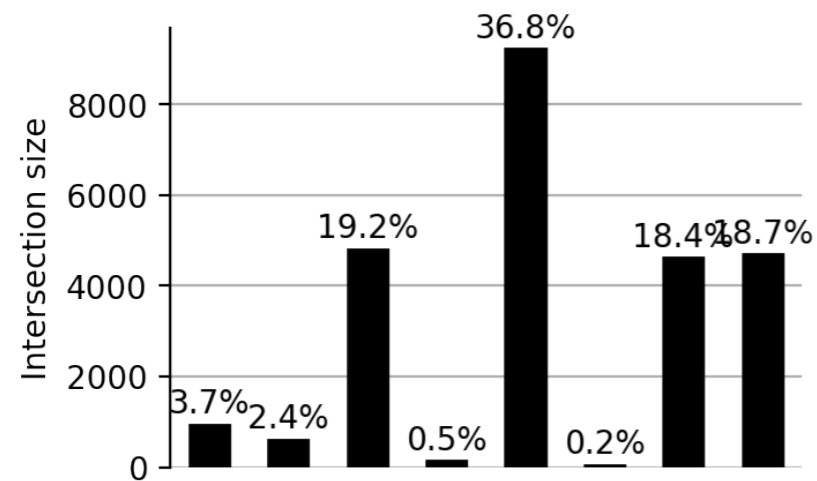


Web tool: <http://fair-checker.france-bioinformatique.fr>

Github: <https://github.com/IFB-ElixirFr/fair-checker>

# Large-scale FAIR metrics evaluations

How FAIR are Bio.Tools registered softwares ?



Running FAIR-Checker over more than **25.000 bioinformatics softwares** from Bio.tools:

R1.1: Finding **licence property**. Only 37,9% of the tools

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

<https://bio.tools/>

# Checking profile conformance ...

[R1.3: \(Meta\)data meet domain-relevant community standards](#)

```
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/...> .
```

## Major issues

This markup is missing  
`dct:conformsTo` properties as well  
as `schema:name` and `schema:url` ...

## Minor issues

This markup should also contains  
`schema:author`, `schema:citation`,  
etc.

Not realistic from a human point of view → automation needed !



# ... to progressively increase metadata completeness

Validation of Bioschemas profiles:

- rank missing metadata
- developer focus first on minimal metadata

*How ? with RDF + SHACL constraints*

The screenshot shows the 'Check BioSchemas' tool interface. At the top, a yellow button labeled 'Check BioSchemas' is centered. Below it, a horizontal line separates the header from the main content. The main content displays the validation of the URL `https://bio.tools/jaspar`, which is identified as a `http://schema.org/SoftwareApplication`. It notes that the validation uses the profile `https://bioschemas.org/profiles/ComputationalTool/1.0-RELEASE` via the `dct:conformsTo` property. Below this, two panels provide details: 'Required missing properties' and 'Improvements'. The 'Required missing properties' panel lists three items that must be provided: `https://schema.org/description`, `https://schema.org/name`, and `https://schema.org/url`. The 'Improvements' panel lists five items that should be provided: `https://schema.org/applicationCategory`, `https://schema.org/author`, `https://schema.org/license`, and `https://schema.org/softwareVersion`.

Check BioSchemas

`https://bio.tools/jaspar` has type `http://schema.org/SoftwareApplication`

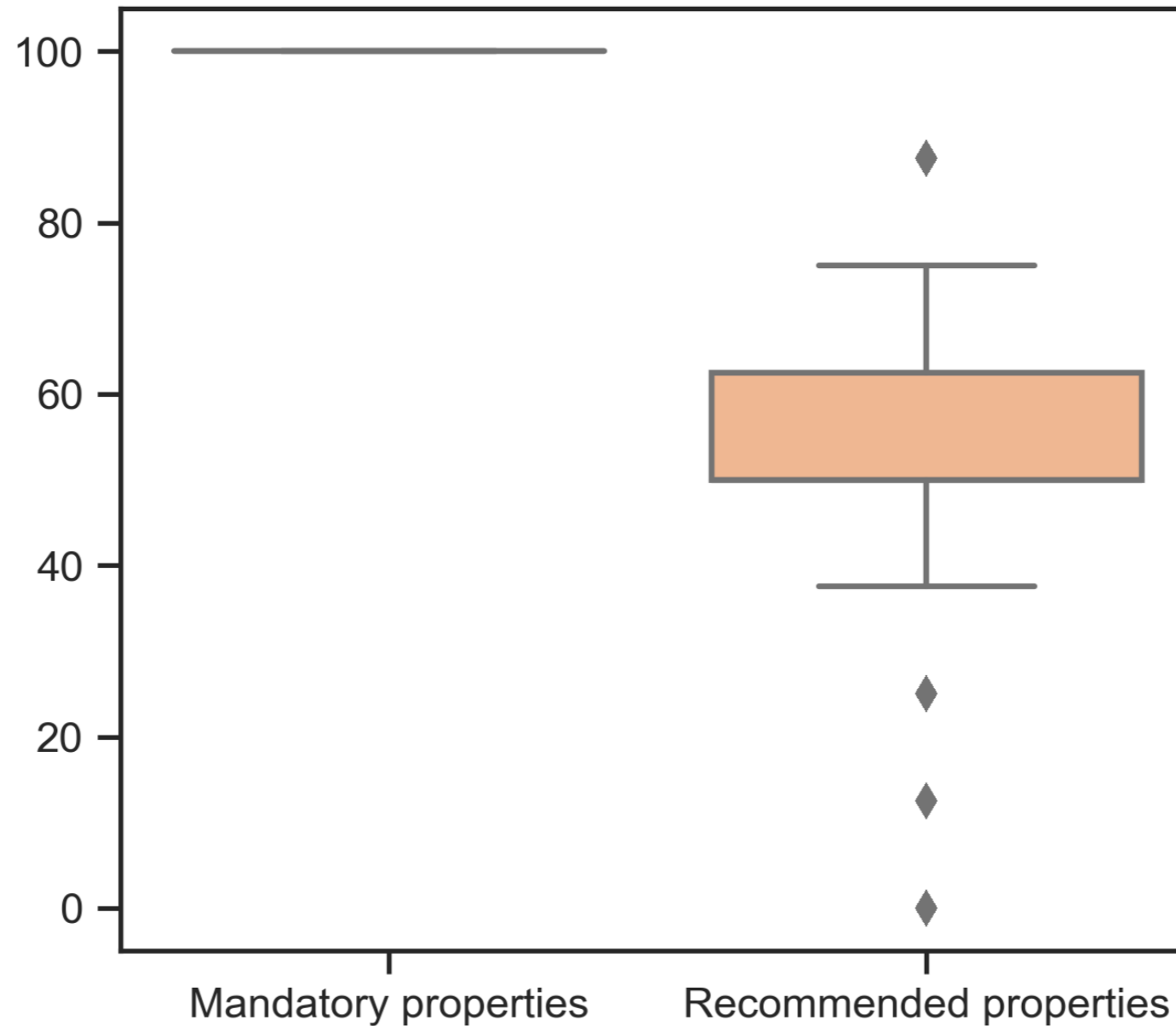
Using `https://bioschemas.org/profiles/ComputationalTool/1.0-RELEASE` for validation, specified from the `dct:conformsTo` property.

Required missing properties	Improvements
<code>https://schema.org/description</code> <b>must be</b> provided	<code>https://schema.org/applicationCategory</code> <b>should be</b> provided
<code>https://schema.org/name</code> <b>must be</b> provided	<code>https://schema.org/author</code> <b>should be</b> provided
<code>https://schema.org/url</code> <b>must be</b> provided	<code>https://schema.org/license</code> <b>should be</b> provided
	<code>https://schema.org/softwareVersion</code> <b>should be</b> provided

# SHACL shapes evaluations

→ machine-actionable bioschemas profile

Compliance of 25048 bioinformatics softwares with the Computational Tool Bioschemas profiles



# Wrap-up & acknowledgments

1. Community agreement on metadata fields to focus on
2. Semantic Search ([schema.org](https://schema.org) + EDAM)
3. Improved FAIRness for software tools



Thomas Rosnet,  
IFB, AMU Marseille



Marie-Dominique Devignes  
Loria, Nancy

Sahar Frikha, Frédéric De Lamotte,  
Vincent Lefort

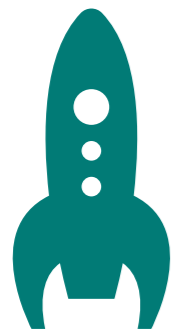
Interoperability Task Force of Elixir-France.



Bioschemas  
community and developers



Bio.Tools developers



More to come this afternoon:  
alignment  $\leftrightarrow$  CodeMeta

Backup slides

# 4. Alignment with other initiatives (CodeMeta)

# CodeMeta

The CodeMeta Project



## Codemeta Terms

### Terms from Schema.org

Recognized properties for CodeMeta **Code** includes the following terms from <https://schema.org>. These terms are part of the CodeMeta specification and can be used without any prefix.

Property	Type	Description
codeRepository	URL	Link to the repository where the un-compiled, human readable code and related code is located (SVN, GitHub, CodePlex, institutional GitLab instance, etc.).
programmingLanguage	ComputerLanguage or Text	The computer programming language.
runtimePlatform	Text	Runtime platform or script interpreter dependencies (Example - Java v1, Python2.3, .Net Framework 3.0). Supersedes runtime.

Dictionary of Schema.org properties relevant for research software

Very similar to Bioschemas

⚠ No priority recommendation (minimum, recommended, optional)

⚠ No recommendation for entity typing: (SoftwareApplication, SoftwareSourceCode)

# Bioschemas & CodeMeta overlap

Bioschemas properties	Marginality	included in CodeMeta ?
schema:description	minimum	Included
schema:name	minimum	Included
schema:url	minimum	Included
schema:applicationCategory	recommended	Included
schema:applicationSubCategory	recommended	Included
schema:author	recommended	Included
schema:citation	recommended	Included
schema:featureList	recommended	Missing
schema:license	recommended	Included
schema:softwareVersion	recommended	Included
schema:applicationSuite	optional	Missing
schema:codeRepository	optional	Included
schema:contributor	optional	Included
schema:discussionUrl	optional	Missing
schema:downloadUrl	optional	Included
schema:funder	optional	Included
schema:hasPart	optional	Included
schema:identifer	optional	Included
<b><i>bioschemas:input</i></b>	optional	Missing
schema:isAccessibleForFree	optional	Included
schema:isBasedOn	optional	Missing
schema:isPartOf	optional	Included
schema:keywords	optional	Included
schema:operatingSystem	optional	Included
<b><i>bioschemas:output</i></b>	optional	Missing
schema:programmingLanguage	optional	Included
schema:provider	optional	Included
schema:softwareAddOn	optional	Missing
schema:softwareHelp	optional	Included
schema:thumbnailUrl	optional	Missing

73% of Bioschemas properties are already in CodeMeta

# CodeMeta not in the Bioschemas (tool profile)

Missing in Bioschemas	Introduced by CodeMeta (not in Schema.org)
runtimePlatform	softwareSuggestions
targetProduct	maintainer
fileSize	contIntegration
installUrl	buildInstructions
memoryRequirements	developmentStatus
permissions	embargoDate
processorRequirements	funding
releaseNotes	issueTracker
softwareRequirements	referencePublication
supportingData	readme
copyrightHolder	
copyrightYear	
creator	
dateCreated	
dateModified	
datePublished	
editor	
encoding	
fileFormat	
producer	
publisher	
sponsor	
version	
position	
sameAs	
relatedLink	
givenName	
familyName	
email	
affiliation	
address	

→ Should we **update** the Computational Tool **profile** with (some of) these properties ?

→ Should we propose **new terms** for Schema.org ?

→ Should we **consume** CodeMeta annotations in the tools ecosystem framework ?

→ Should we **publish** bio.tools content in CodeMeta compatible registry ?



# SHACL

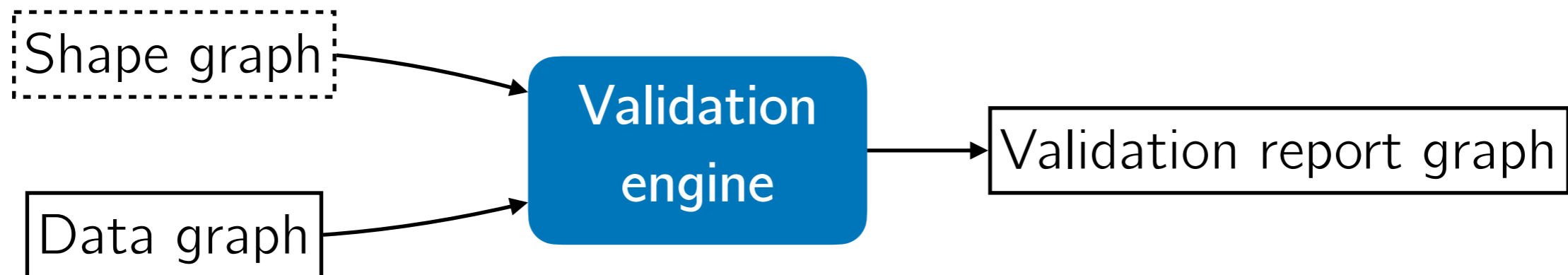
Stands for "SHApes Constraint Language".

W3C recommendation (July 2017) aimed at validating RDF graphs.

Similar to the ShEx (Shape Expressions) initiative.

Shape = pattern / constraints for an RDF graph

SHACL shapes are written with RDF triples



# Validation report

Validation Report

**Conforms: False**

Results (2):

Constraint Violation in MinCountConstraintComponent  
(<http://www.w3.org/ns/shacl#MinCountConstraintComponent>):

Severity: **sh:Violation**

Source Shape: [ sh:minCount Literal("1",  
datatype=xsd:integer) ; sh:path sc:name ; sh:severity  
sh:Violation ]

Focus Node: ex:myTool

Result Path: sc:name

Message: **Less than 1 values on ex:myTool->sc:name**

Validation Result in MinCountConstraintComponent (<http://www.w3.org/ns/shacl#MinCountConstraintComponent>):

Severity: **sh:Warning**

Source Shape: [ sh:minCount Literal("1",  
datatype=xsd:integer) ; sh:path sc:citation ; sh:severity  
sh:Warning ]

Focus Node: ex:myTool

Result Path: sc:citation

Message: **Less than 1 values on ex:myTool->sc:citation**

Depending on the evaluation engine, you can get a textual report:

- ▶ Yes/No answer for the global validation
- ▶ One message per error
- ▶ Source shape leading to error
- ▶ Focus node leading to error

The report is generated from the validation report graph.

# Validation report

```
@prefix sc: <http://schema.org/> .
@prefix sh: <http://www.w3.org/ns/shacl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

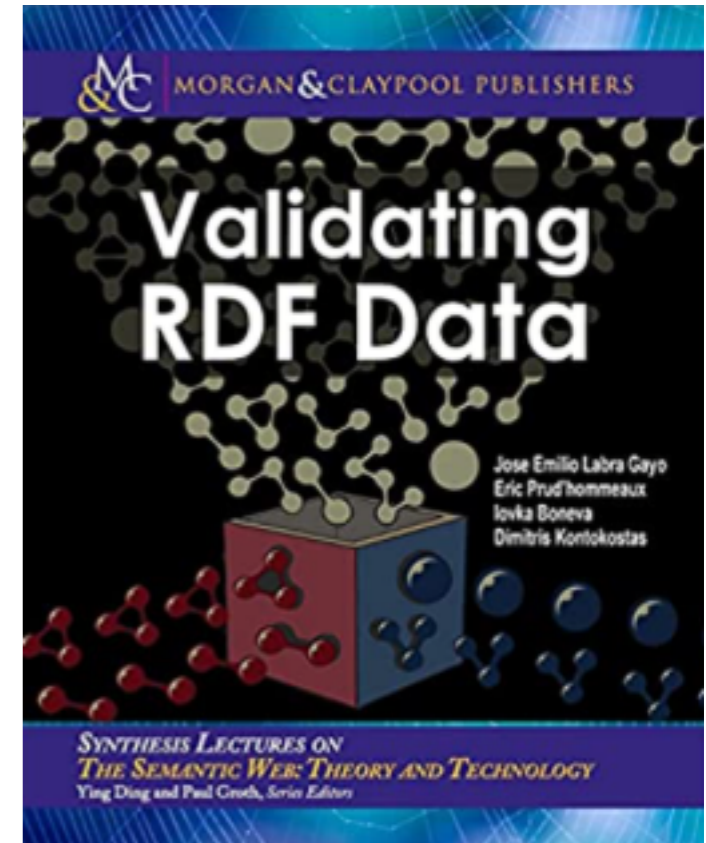
[] a sh:ValidationReport ;
  sh:conforms false ;
  sh:result [ a sh:ValidationResult ;
    sh:focusNode <http://
bioschemas.validation.tutorial/myTool> ;
    sh:resultMessage "Less than 1 values on
ex:myTool->sc:citation" ;
    sh:resultPath sc:citation ;
    sh:resultSeverity sh:Warning ;
    sh:sourceConstraintComponent
sh:MinCountConstraintComponent ;
    sh:sourceShape [ sh:minCount 1 ;
      sh:path sc:citation ;
      sh:severity sh:Warning ] ],
  [ a sh:ValidationResult ;
    sh:focusNode <http://
bioschemas.validation.tutorial/myTool> ;
    sh:resultMessage "Less than 1 values on
ex:myTool->sc:name" ;
    sh:resultPath sc:name ;
    sh:resultSeverity sh:Violation ;
    sh:sourceConstraintComponent
sh:MinCountConstraintComponent ;
    sh:sourceShape [ sh:minCount 1 ;
      sh:path sc:name ;
      sh:severity sh:Violation ] ] .
```

- ▶ SHACL provides a controlled vocabulary to describe validation reports in RDF.
- ▶ Validation reports can be shared and queried on the web following Linked Data principles.

# To go further ...

José Emilio Labra Gayo, Eric Prud'hommeaux, Iovka Boneva and Dimitris Kontokostas. "Validating RDF Data." *Validating RDF Data* (2017).

Online version: <https://book.validatingrdf.com>



ISWC 2020 tutorial, Jose Emilio Labra Gayo:

<http://www.validatingrdf.com/tutorial/iswc2020/>