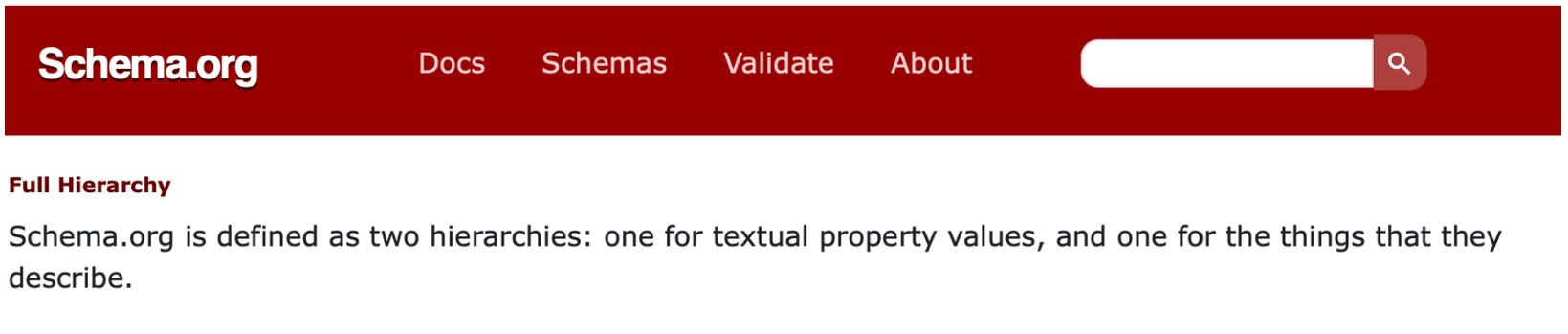
(more) Findable bionformatics softwares with Bioschemas

Elixir Tools Platform F2F meeting, Leiden, 28/03/2023

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

Bioschemas

schema.org



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

Schema.org is massively adopted

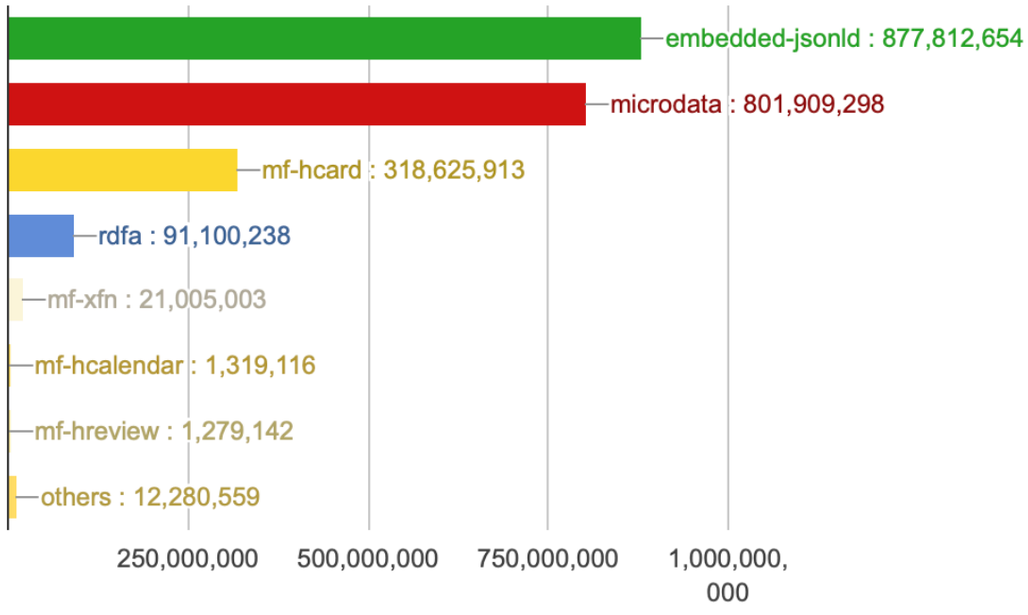
Web Data Commons

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

Schema.org for Life-Science resources ?

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					

- 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

```
ex:myTool    rdf:type    schema:SoftwareApplication .
```

Example

[...]

Property	Expected Type	Description	CD	Controlled Vocabulary	Example
Marginality: Minimum.					
@context	URL	Used to provide the context (namespaces) for the JSON-LD file. Not needed in other serialisations.	ONE		
@type	Text	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	
@id	IRI	Used to distinguish the resource being described in JSON-LD. For other serialisations use the appropriate approach.	ONE		
dct:conformsTo	IRI	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 (http://purl.org/dc/terms/conformsTo), see example.	ONE	Bioschemas profile versioned URL	
description	Text	**Schema:** A description of the item. **Bioschemas:** A short description of the tool.	ONE		

[...]

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

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

- **Instrumented** bio.tools **registry** to produce Bioschemas markup
  → cost of annotating a software = cost of publishing it through bio.tools
- Tools ecosystem
  → **automated** feeding of Bioschemas **data dump** (GitHub actions)
- **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

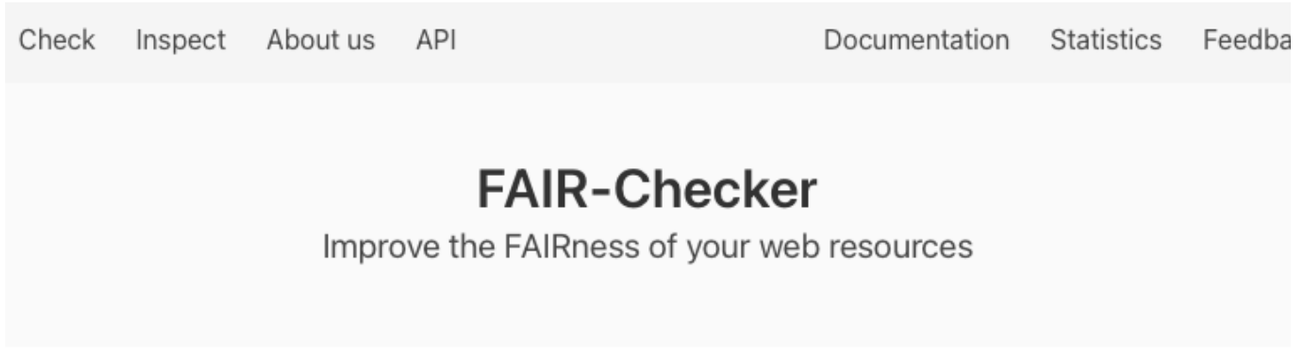
≠ technical specifications

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

Objectives

(i) monitor the results of FAIR metrics evaluation

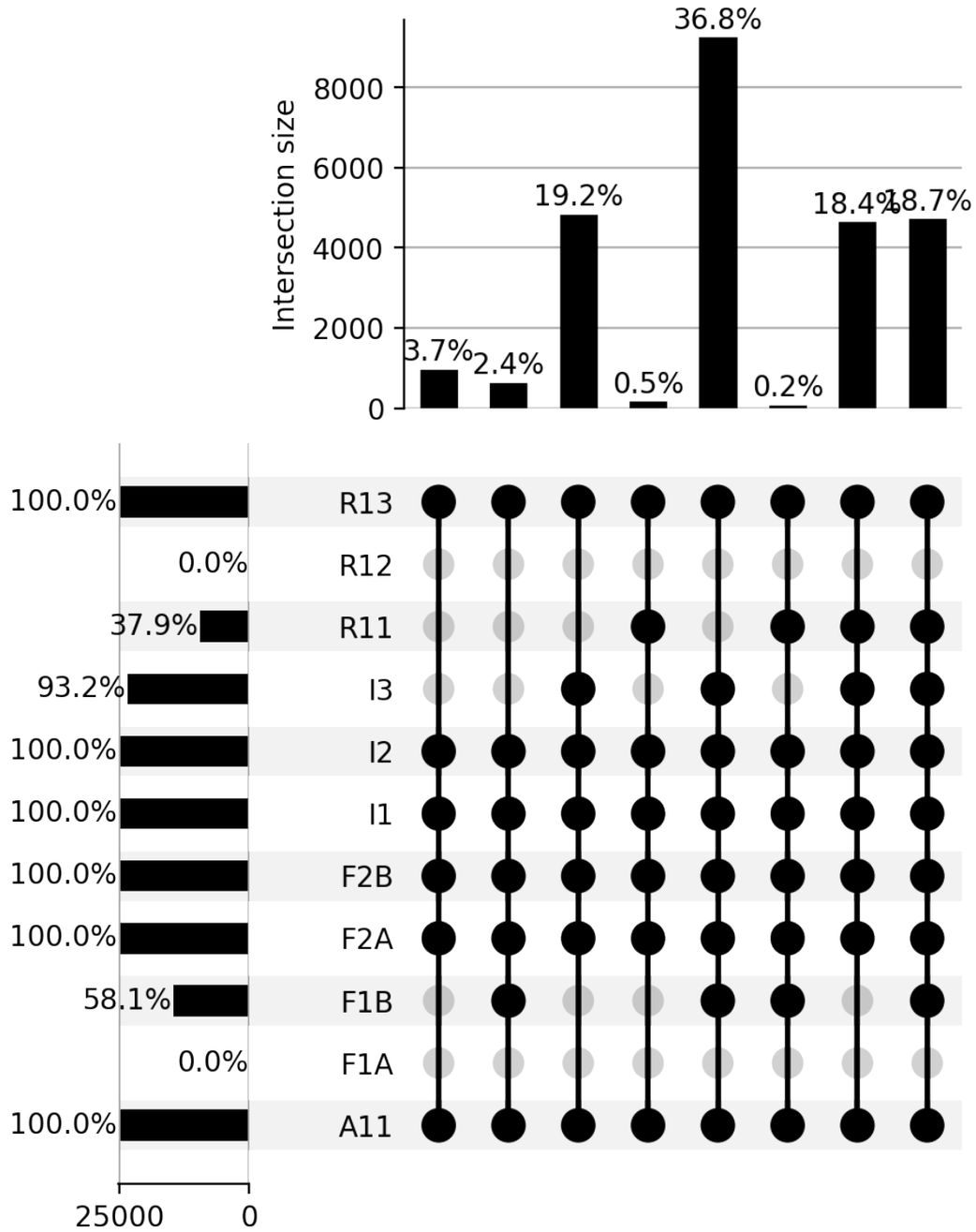
(ii) improve the quality of embedded metadata



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 ?



https://bio.tools/

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

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.

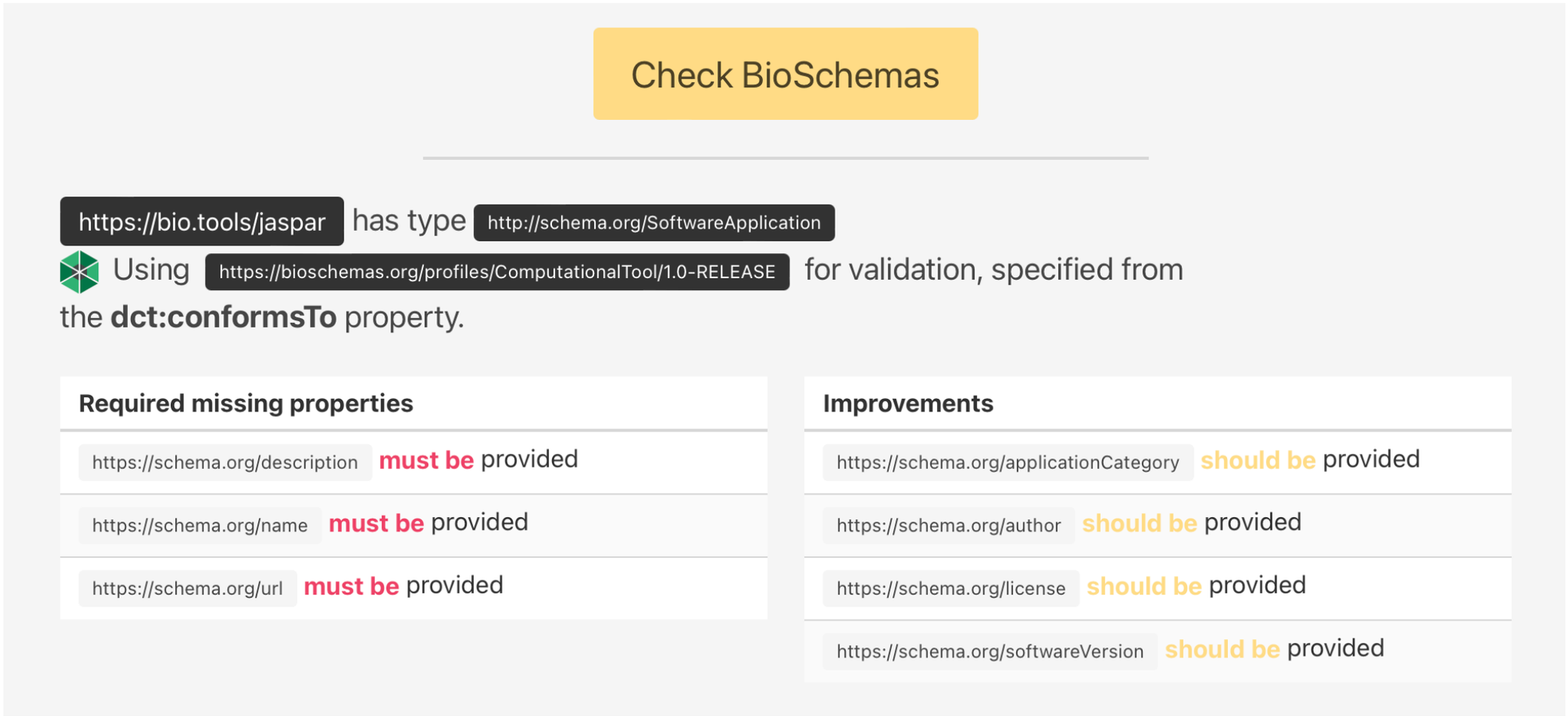
Not realistic from a human point of view $\rightarrow$ automation needed !

... to progressively increase metadata completeness

Validation of Bioschemas profiles:

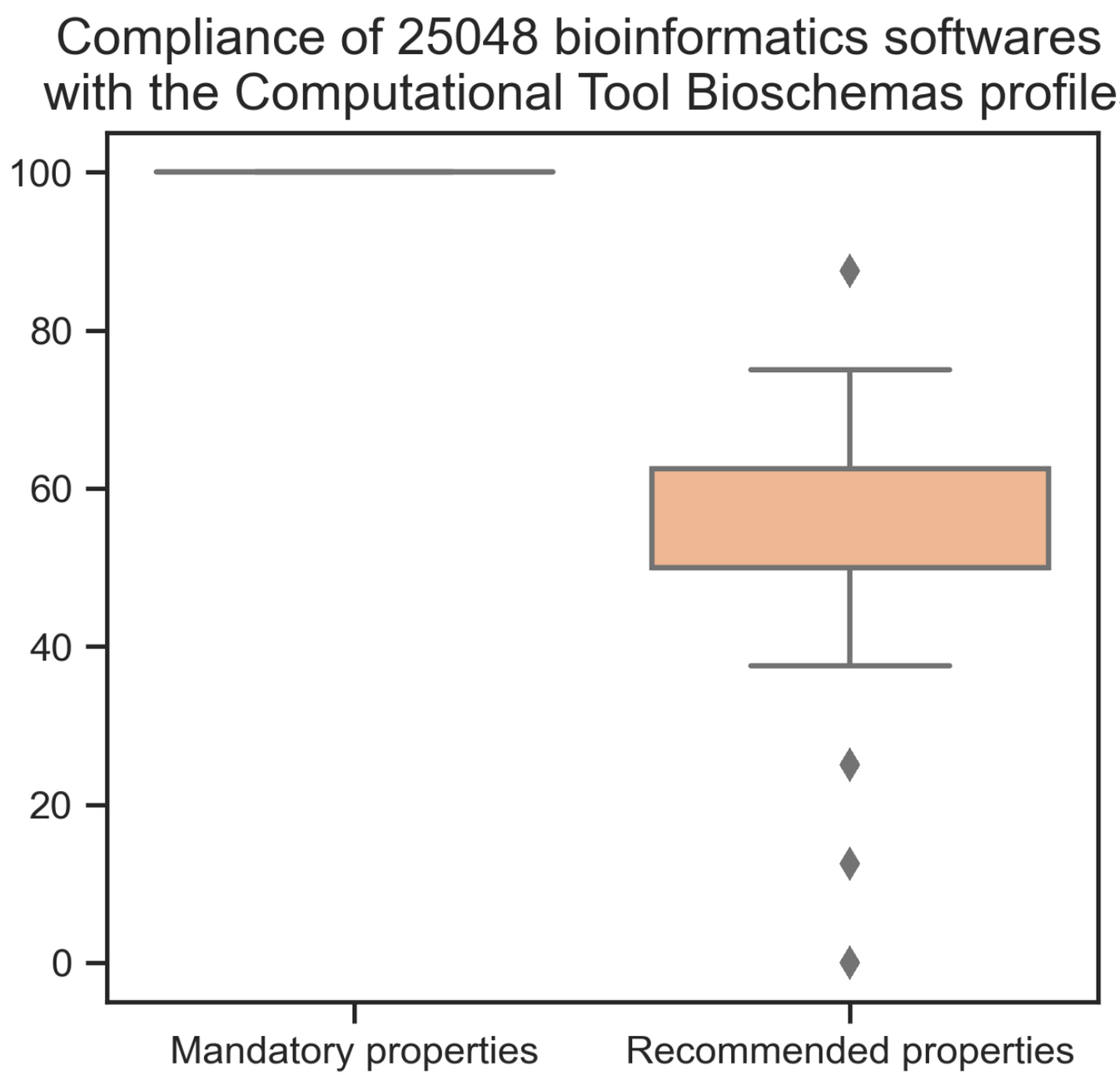
→ **rank missing** metadata
→ developer **focus** first
on **minimal** metadata

How ? with RDF + SHACL constraints



SHACL shapes evaluations

→ machine-actionable bioschemas profile



4. Alignment with other initatives (CodeMeta)

CodeMeta

The CodeMeta Project

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.

Dictionnary of Shema.org properties relevant for research software

Very similar to Bioschemas

No priority recommendation (minimum, recommened, 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
bioschemas:input	optional	Missing
schema:isAccessibleForFree	optional	Included
schema:isBasedOn	optional	Missing
schema:isPartOf	optional	Included
schema:keywords	optional	Included
schema:operatingSystem	optional	Included
bioschemas:output	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 ?

Wrap-up & acknowledgments

1. Community agreement on metadata fields to focus onVery similar to Bioschemas
2. Semantic Search (schema.org + EDAM)
3. Improved FAIRness for tools
4. Aligned with other FAIR initiatives (CodeMeta)

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