FAIR Principles practices supporting recognition

(experiences from ELIXIR and FAIRDOM)

Professor Carole Goble

The University of Manchester, UK

elixir

Head of UK Node
Co-lead Interoperability
Platform

FAIRDOM

Co-ordinator
Chair FAIRDOM
Association e.V.

Bioschemas.org

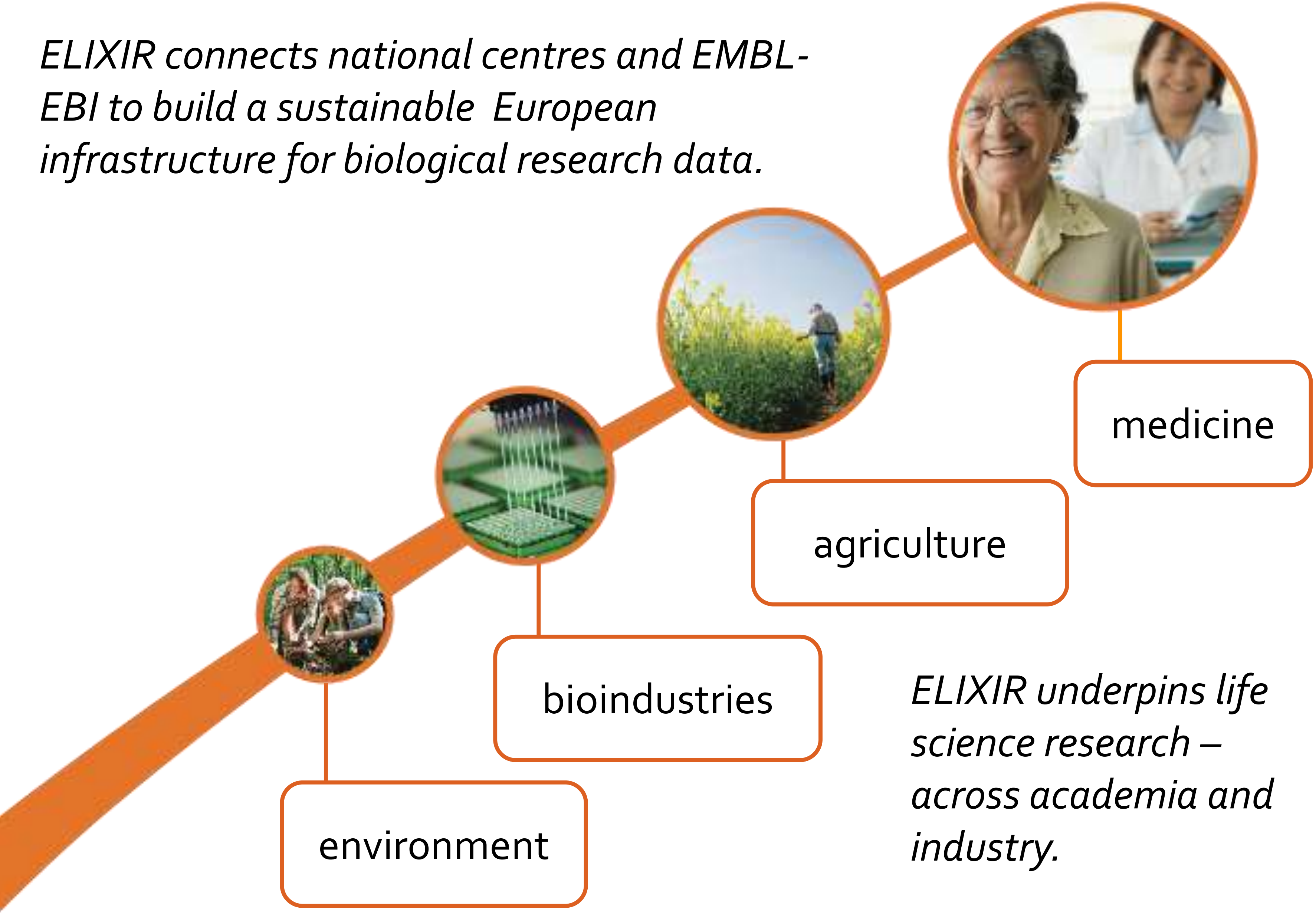
Leadership team

FORCE11
The Future of Research Communications and e-Scholarship

DCIP Advisor
Advisory Board

UK Biobanking Showcase 2017, London, 18 October 2017

ELIXIR connects national centres and EMBL-EBI to build a sustainable European infrastructure for biological research data.



environment

bioindustries

agriculture

medicine

ELIXIR underpins life science research – across academia and industry.

European Research Infrastructure

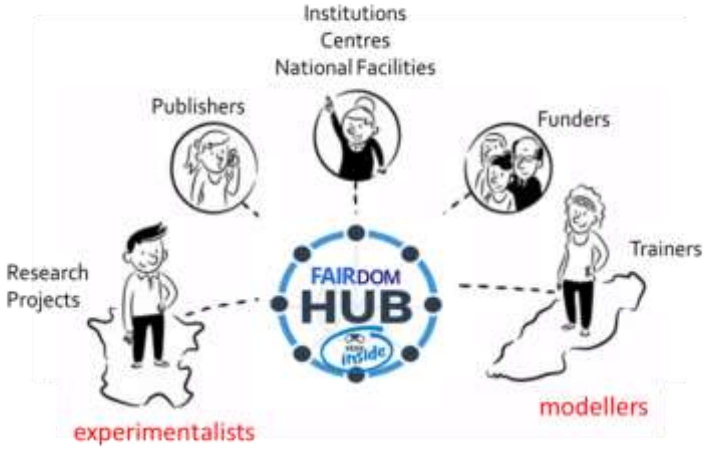
Life Sciences Data
Standards, Portals, Platforms, Lobbying, Sustain Core Resources
Archive Level: e.g. BioSamples

Sweden
Belarus
Ukraine
Romania
Greece
Turkey
Tunisia
Morocco
Algeria

22 Countries

Systems and Synthetic Biology Projects

Management for Research Data, Operations, Models Assets at the Project level
e.g. Samples, Strains, Specimens



80+ Projects
700+ Researchers

Recognition: being FAIR

Find – Download – Go

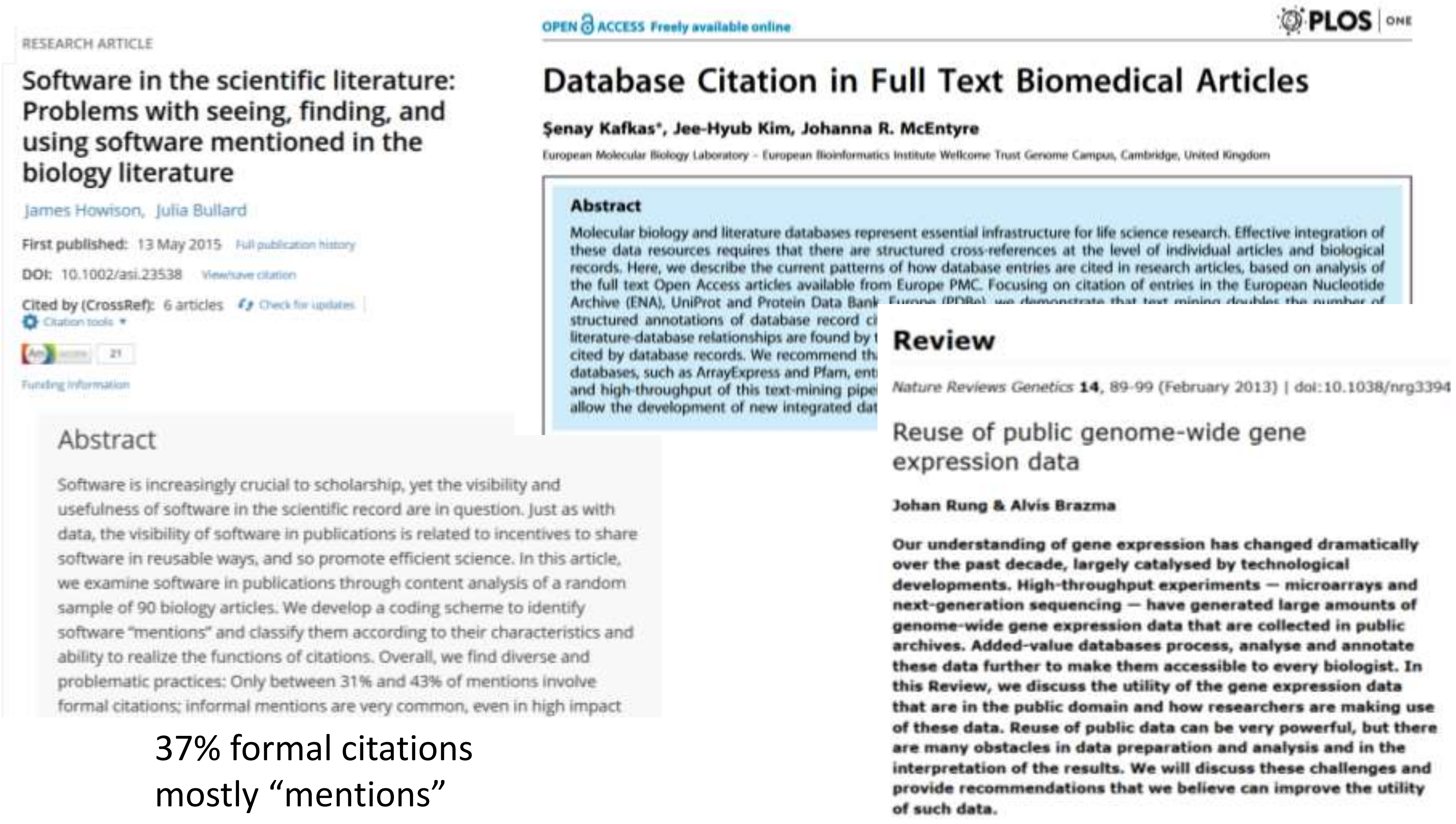
Data, model, SOP, sample… provider

make it easier to be found and to track credit

Data, model, SOP, sample… user

make it easier to find and to action credit

Getting recognition



RESEARCH ARTICLE

Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature

James Howison, Julia Bullard

First published: 13 May 2015 Full publication history

DOI: 10.1002/asi.23538 View/save citation

Cited by (CrossRef): 6 articles Check for updates

Citation tools ▾

21

Funding information

Abstract

Software is increasingly crucial to scholarship, yet the visibility and usefulness of software in the scientific record are in question. Just as with data, the visibility of software in publications is related to incentives to share software in reusable ways, and so promote efficient science. In this article, we examine software in publications through content analysis of a random sample of 90 biology articles. We develop a coding scheme to identify software "mentions" and classify them according to their characteristics and ability to realize the functions of citations. Overall, we find diverse and problematic practices: Only between 31% and 43% of mentions involve formal citations; informal mentions are very common, even in high impact

OPEN ACCESS Freely available online

PLOS | ONE

Database Citation in Full Text Biomedical Articles

Şenay Kafkas*, Jee-Hyub Kim, Johanna R. McEntyre

European Molecular Biology Laboratory - European Bioinformatics Institute Wellcome Trust Genome Campus, Cambridge, United Kingdom

Abstract

Molecular biology and literature databases represent essential infrastructure for life science research. Effective integration of these data resources requires that there are structured cross-references at the level of individual articles and biological records. Here, we describe the current patterns of how database entries are cited in research articles, based on analysis of the full text Open Access articles available from Europe PMC. Focusing on citation of entries in the European Nucleotide Archive (ENA), UniProt and Protein Data Bank Europe (PDBe), we demonstrate that text mining doubles the number of structured annotations of database record ci
literature-database relationships are found by t
cited by database records. We recommend th
databases, such as ArrayExpress and Pfam, ent
and high-throughput of this text-mining pipe
allow the development of new integrated dat

Review

Nature Reviews Genetics **14**, 89-99 (February 2013) | doi:10.1038/nrg3394

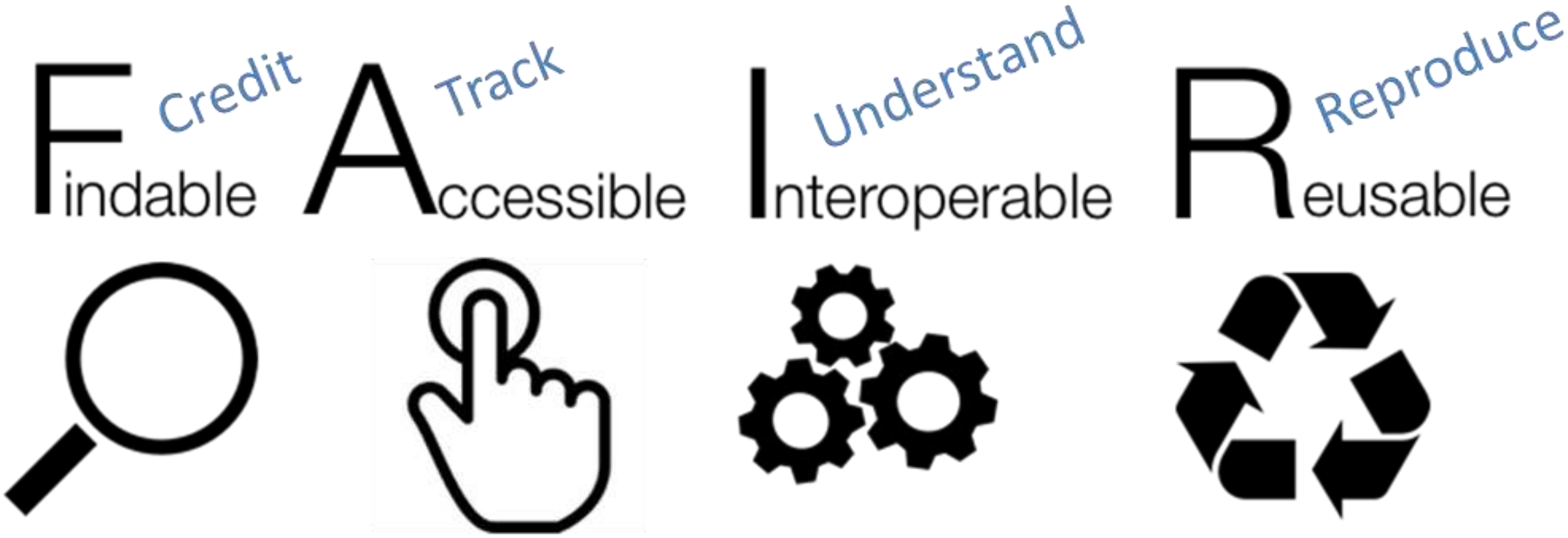
Reuse of public genome-wide gene expression data

Johan Rung & Alvis Brazma

Our understanding of gene expression has changed dramatically over the past decade, largely catalysed by technological developments. High-throughput experiments — microarrays and next-generation sequencing — have generated large amounts of genome-wide gene expression data that are collected in public archives. Added-value databases process, analyse and annotate these data further to make them accessible to every biologist. In this Review, we discuss the utility of the gene expression data that are in the public domain and how researchers are making use of these data. Reuse of public data can be very powerful, but there are many obstacles in data preparation and analysis and in the interpretation of the results. We will discuss these challenges and provide recommendations that we believe can improve the utility of such data.

37% formal citations
mostly "mentions"

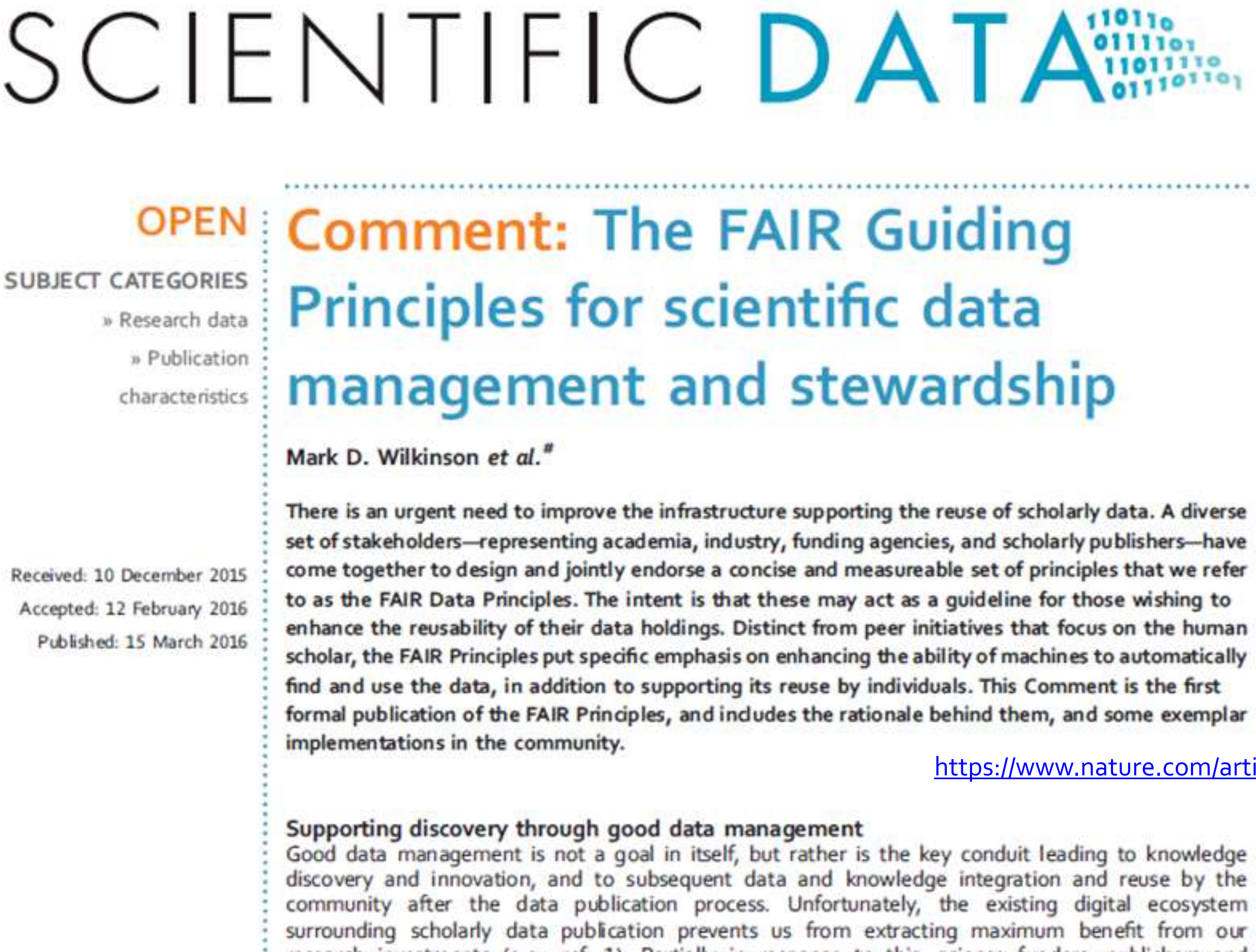
25% Publications that used the public Arrayexpress Archive cited it



The FAIR Guiding Principles for scientific data management and stewardship
https://www.nature.com/articles/sdata201618 (2016)

FAIR Data Principles

Access to public funded research, Reproducible results
Value and CREDIT all research outputs



SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

https://www.nature.com/articles/sdata201618 (2016)

Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our

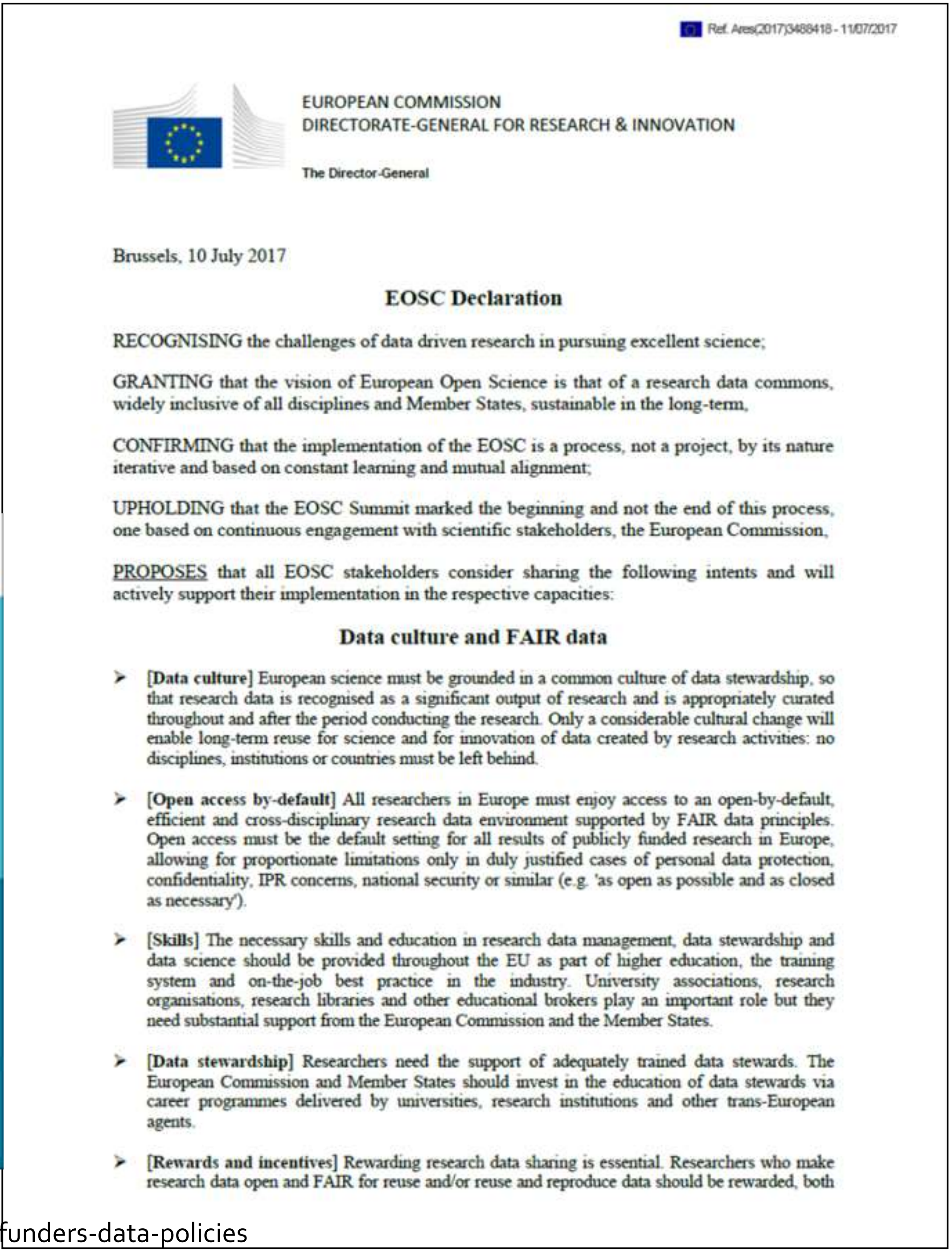
EUROPEAN COMMISSION
Directorate-General for Research & Innovation

H2020 Programme

Guidelines on
FAIR Data Management in Horizon 2020

Realising
the European
Open Science Cloud

First report and recommendations
of the Commission High Level Expert Group
on the European Open Science Cloud



Ref. Ares(2017)3488418 - 11/07/2017

EUROPEAN COMMISSION
DIRECTORATE-GENERAL FOR RESEARCH & INNOVATION

The Director-General

Brussels, 10 July 2017

EOSC Declaration

RECOGNISING the challenges of data driven research in pursuing excellent science;

GRANTING that the vision of European Open Science is that of a research data commons, widely inclusive of all disciplines and Member States, sustainable in the long-term,

CONFIRMING that the implementation of the EOSC is a process, not a project, by its nature iterative and based on constant learning and mutual alignment;

UPHOLDING that the EOSC Summit marked the beginning and not the end of this process, one based on continuous engagement with scientific stakeholders, the European Commission,

PROPOSES that all EOSC stakeholders consider sharing the following intents and will actively support their implementation in the respective capacities:

Data culture and FAIR data

➢ **[Data culture]** European science must be grounded in a common culture of data stewardship, so that research data is recognised as a significant output of research and is appropriately curated throughout and after the period conducting the research. Only a considerable cultural change will enable long-term reuse for science and for innovation of data created by research activities: no disciplines, institutions or countries must be left behind.

➢ **[Open access by-default]** All researchers in Europe must enjoy access to an open-by-default, efficient and cross-disciplinary research data environment supported by FAIR data principles. Open access must be the default setting for all results of publicly funded research in Europe, allowing for proportionate limitations only in duly justified cases of personal data protection, confidentiality, IPR concerns, national security or similar (e.g. 'as open as possible and as closed as necessary').

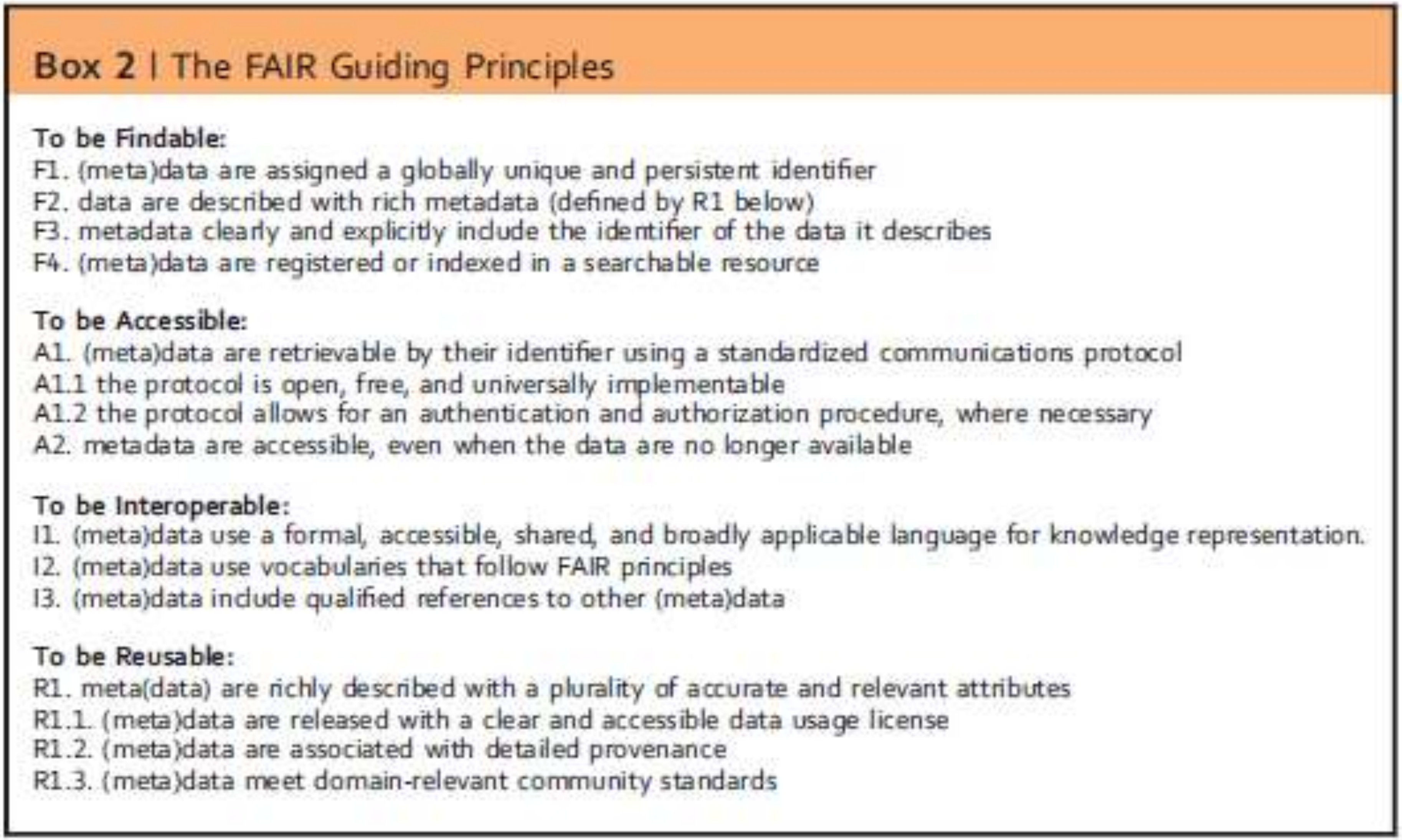
➢ **[Skills]** The necessary skills and education in research data management, data stewardship and data science should be provided throughout the EU as part of higher education, the training system and on-the-job best practice in the industry. University associations, research organisations, research libraries and other educational brokers play an important role but they need substantial support from the European Commission and the Member States.

➢ **[Data stewardship]** Researchers need the support of adequately trained data stewards. The European Commission and Member States should invest in the education of data stewards via career programmes delivered by universities, research institutions and other trans-European agents.

➢ **[Rewards and incentives]** Rewarding research data sharing is essential. Researchers who make research data open and FAIR for reuse and/or reuse and reproduce data should be rewarded, both

UK Funder Data Policies
http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies

Machine Processability for Distributed Systems



Box 2 | The FAIR Guiding Principles

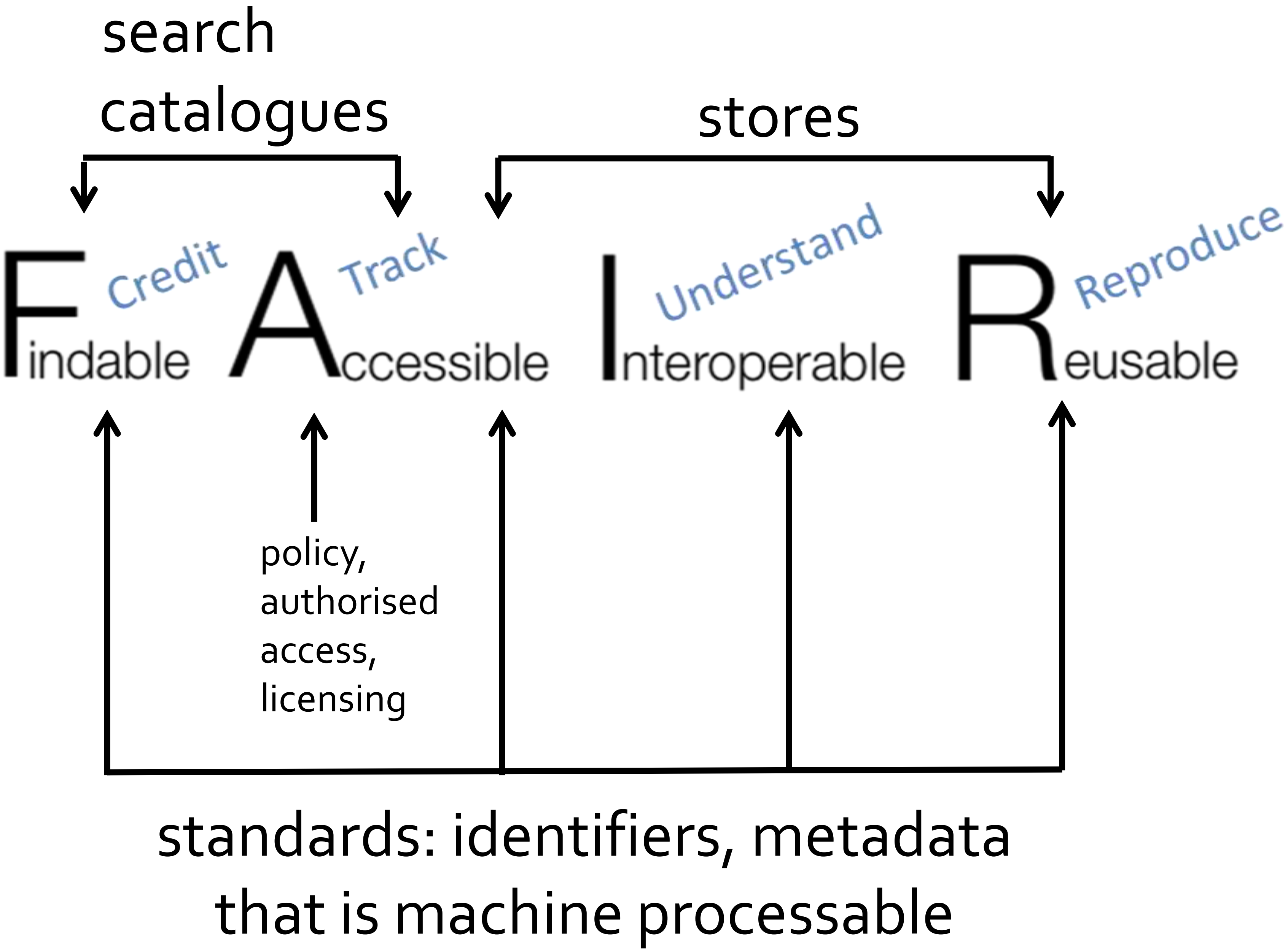
To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

The FAIR Guiding Principles for scientific data management and stewardship
https://www.nature.com/articles/sdata201618 (2016)



search catalogues

stores

Findable Credit

Accessible Track

Interoperable Understand

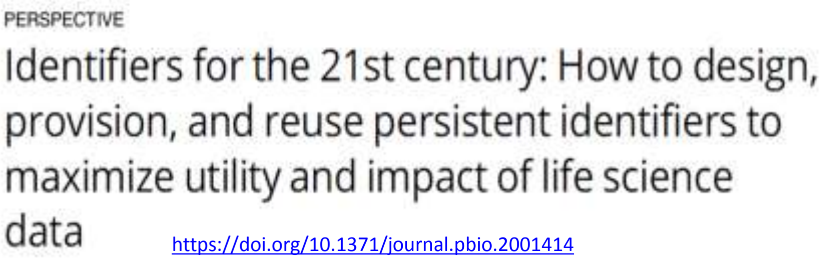
Reusable Reproduce

policy, authorised access, licensing

standards: identifiers, metadata that is machine processable

1. Identifiers and Citation

Best Practices & Principles



PERSPECTIVE

Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data

https://doi.org/10.1371/journal.pbio.2001414

#CiteTheData campaigns

DC1

Data Citation Principles

<JATS>

https://www.force11.org/group/joint-declaration-data-citation-principles-final

Data Citation Implementation

Identifier Schemes

Resource Identification Initiative

doi

https://www.force11.org/group/resource-identification-initiative

Identifiers.org

DataCite

FIND, ACCESS, AND REUSE DATA

Impactstory

Services

Identifiers for 21st Century

1. Credit any derived content using its original identifier
2. Help local IDs travel well: Document prefix and patterns
3. Opt for simple, durable web resolution
4. Avoid embedding meaning or relying on it for uniqueness
5. Design new identifiers for diverse uses by others
6. Implement a version-management policy
7. Do not reassign or delete identifiers
8. **Make URIs clear and findable**
9. Document the identifiers you issue and use
10. Reference and display responsibly

Data Citation Principles

- 1. Importance
- **2. Credit and Attribution** Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data
- 3. Evidence
- **4. Unique Identification** A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.
- **5. Access** Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.
- 6. Persistence
- 7. Specificity and Verifiability
- 8. Interoperability and Flexibility

2. Find and Access

schema.org

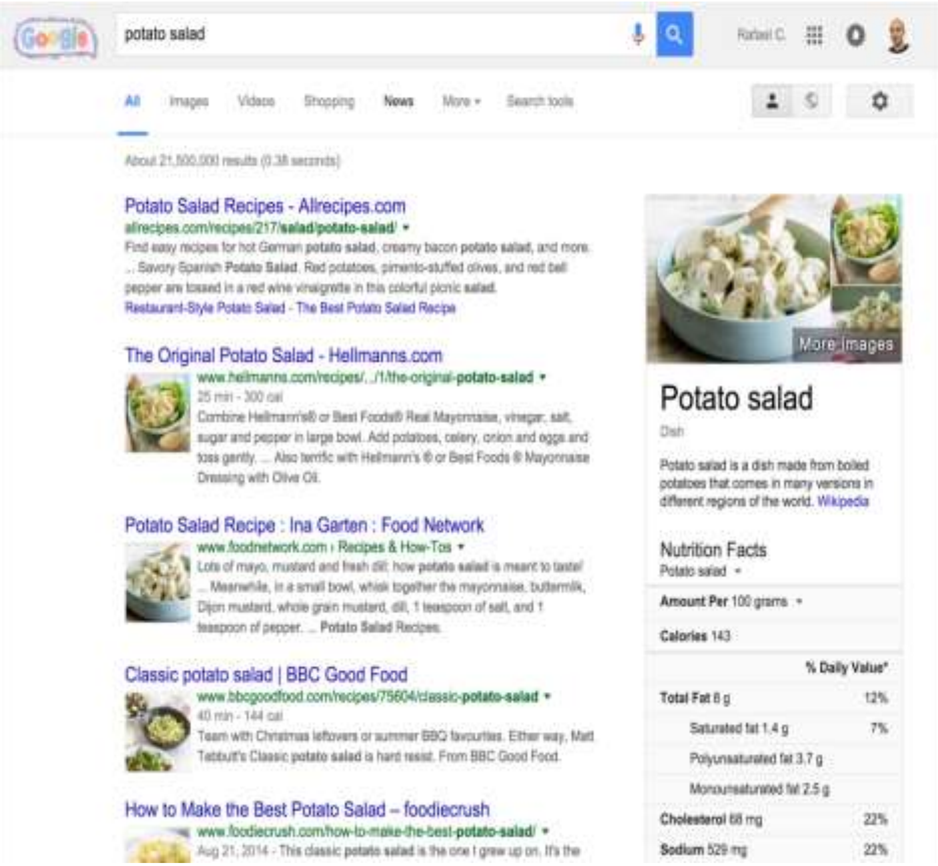
Google YAHOO! Yandex bing

Schema.org adds simple **structured metadata markup** to web pages & sitemaps for harvesting, search and summary snippet making.

Search engines often highlight websites containing Schema.org

Widespread commercial and open source infrastructure -> low barrier to adoption.

No need for an API or special feeds.



About 21,500,000 results (0.38 seconds)

Potato Salad Recipes - Allrecipes.com
allrecipes.com/recipes/217/**salad**/**potato-salad**/ ▾
Find easy recipes for hot German **potato salad**, creamy bacon **potato salad**, and more. ... Savory Spanish **Potato Salad**. Red potatoes, pimento-stuffed olives, and red bell pepper are tossed in a red wine vinaigrette in this colorful picnic **salad**.
Restaurant-Style Potato Salad - The Best Potato Salad Recipe

The Original Potato Salad - Hellmanns.com
www.hellmanns.com/recipes/.../1/the-original-**potato-salad** ▾
25 min - 300 cal
Combine Hellmann's® or Best Foods® Real Mayonnaise, vinegar, salt, sugar and pepper in large bowl. Add potatoes, celery, onion and eggs and toss gently. ... Also terrific with Hellmann's ® or Best Foods ® Mayonnaise Dressing with Olive Oil.

Potato Salad Recipe : Ina Garten : Food Network
www.foodnetwork.com › Recipes & How-Tos ▾
Lots of mayo, mustard and fresh dill: how **potato salad** is meant to taste! ... Meanwhile, in a small bowl, whisk together the mayonnaise, buttermilk, Dijon mustard, whole grain mustard, dill, 1 teaspoon of salt, and 1 teaspoon of pepper. ... **Potato Salad** Recipes.

Classic potato salad | BBC Good Food
www.bbcgoodfood.com/recipes/75604/classic-**potato-salad** ▾
40 min - 144 cal
Team with Christmas leftovers or summer BBQ favourites. Either way, Matt Tebbutt's Classic **potato salad** is hard resist. From BBC Good Food.

How to Make the Best Potato Salad – foodiecrush
www.foodiecrush.com/how-to-make-the-best-**potato-salad**/ ▾
Aug 21, 2014 - This classic **potato salad** is the one I grew up on. It's the

Potato salad
Dish

Potato salad is a dish made from boiled potatoes that comes in many versions in different regions of the world. Wikipedia

Nutrition Facts
Potato salad ▾

Amount Per 100 grams ▾

Calories 143

	% Daily Value*
Total Fat 8 g	12%
Saturated fat 1.4 g	7%
Polyunsaturated fat 3.7 g	
Monounsaturated fat 2.5 g	
Cholesterol 68 mg	22%
Sodium 529 mg	22%

Bioschemas.org

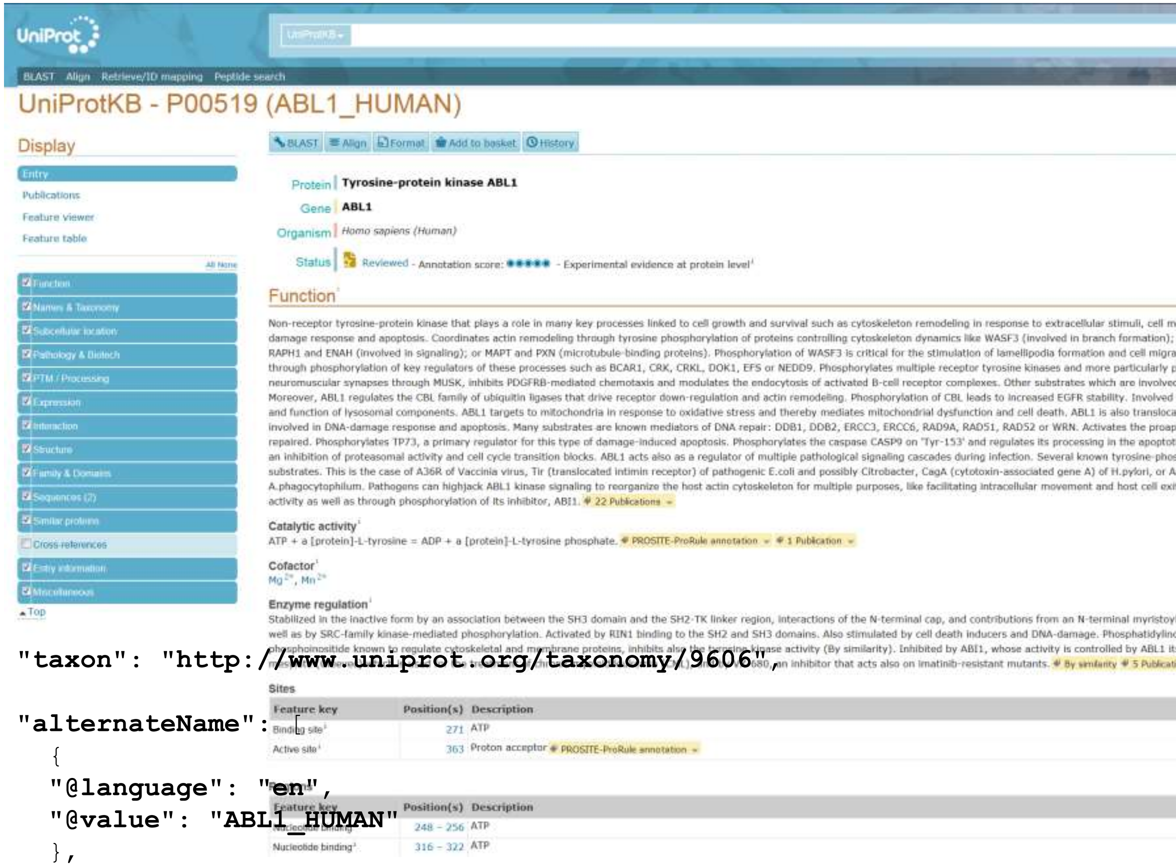
schema.org tailored to the Biosciences

From Potato Salad to Protein Annotation

```
{
"@context": "http://schema.org",
"@type": "BiologicalEntity",
"@id":
"http://www.identifiers.org/uniprot/P00519
",

"biologicalType": "protein",
"isMentionedIn": {
  "@type": "Dataset",
  "@id":
"http://www.uniprot.org/news/2017/03/15/re
lease"
},

"associatedDisease": {
  "@type": "MedicalCondition",
  "@id":
"http://www.omim.org/entry/608232",
  "name": "Leukemia, chronic myeloid
(CML)",
  "code": {
    "@type": "MedicalCode",
    "code": "608232",
    "codingSystem": "OMIM"
  },
  "sameAs":
"http://www.uniprot.org/diseases/DI-03735"
},
"biocoordinates": {
  "@type": "QuantitativeValue",
  "value": "1130"
},
```



UniProt

UniProtKB

BLAST Align Retrieve/ID mapping Peptide search

UniProtKB - P00519 (ABL1_HUMAN)

Display

Entry

Publications

Feature viewer

Feature table

All None

☑ Function
☑ Names & Taxonomy
☑ Subcellular location
☑ Pathology & Biotech
☑ PTM / Processing
☑ Expression
☑ Interaction
☑ Structure
☑ Family & Domains
☑ Sequences (2)
☑ Similar proteins
□ Cross-references
☑ Entry information
☑ Miscellaneous

Top

BLAST Align Format Add to basket History

Protein | Tyrosine-protein kinase ABL1

Gene | ABL1

Organism | *Homo sapiens (Human)*

Status | Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level

Function

Non-receptor tyrosine-protein kinase that plays a role in many key processes linked to cell growth and survival such as cytoskeleton remodeling in response to extracellular stimuli, cell m
damage response and apoptosis. Coordinates actin remodeling through tyrosine phosphorylation of proteins controlling cytoskeleton dynamics like WASF3 (involved in branch formation);
RAPH1 and ENAH (involved in signaling); or MAPT and PXN (microtubule-binding proteins). Phosphorylation of WASF3 is critical for the stimulation of lamellipodia formation and cell migra
through phosphorylation of key regulators of these processes such as BCAR1, CRK, CRKL, DOK1, EFS or NEDD9. Phosphorylates multiple receptor tyrosine kinases and more particularly p
neuromuscular synapses through MUSK, inhibits PDGFRB-mediated chemotaxis and modulates the endocytosis of activated B-cell receptor complexes. Other substrates which are involved
Moreover, ABL1 regulates the CBL family of ubiquitin ligases that drive receptor down-regulation and actin remodeling. Phosphorylation of CBL leads to increased EGFR stability. Involved
and function of lysosomal components. ABL1 targets to mitochondria in response to oxidative stress and thereby mediates mitochondrial dysfunction and cell death. ABL1 is also transloca
involved in DNA-damage response and apoptosis. Many substrates are known mediators of DNA repair: DDB1, DDB2, ERCC3, ERCC6, RAD9A, RAD51, RAD52 or WRN. Activates the proap
repaired. Phosphorylates TP73, a primary regulator for this type of damage-induced apoptosis. Phosphorylates the caspase CASP9 on 'Tyr-153' and regulates its processing in the apoptot
an inhibition of proteasomal activity and cell cycle transition blocks. ABL1 acts also as a regulator of multiple pathological signaling cascades during infection. Several known tyrosine-phos
substrates. This is the case of A36R of Vaccinia virus, Tir (translocated intimin receptor) of pathogenic E.coli and possibly Citrobacter, CagA (cytotoxin-associated gene A) of H.pylori, or A
A.phagocytophilum. Pathogens can highjack ABL1 kinase signaling to reorganize the host actin cytoskeleton for multiple purposes, like facilitating intracellular movement and host cell exi
activity as well as through phosphorylation of its inhibitor, ABI1. 22 Publications

Catalytic activity

ATP + a [protein]-L-tyrosine = ADP + a [protein]-L-tyrosine phosphate. PROSITE-ProRule annotation 1 Publication

Cofactor

Mg^{2+}, Mn^{2+}

Enzyme regulation

Stabilized in the inactive form by an association between the SH3 domain and the SH2-TK linker region, interactions of the N-terminal cap, and contributions from an N-terminal myristoy
well as by SRC-family kinase-mediated phosphorylation. Activated by RIN1 binding to the SH2 and SH3 domains. Also stimulated by cell death inducers and DNA-damage. Phosphatidylino
phosphoinositide known to regulate cytoskeletal and membrane proteins, inhibits also the tyrosine kinase activity (By similarity). Inhibited by ABI1, whose activity is controlled by ABL1 it
an inhibitor that acts also on imatinib-resistant mutants. By similarity 5 Publicati

Sites

Feature key	Position(s)	Description
Binding site	271	ATP
Active site	363	Proton acceptor PROSITE-ProRule annotation

Feature key	Position(s)	Description
Nucleotide binding	248 - 256	ATP
Nucleotide binding	316 - 322	ATP

```
"taxon": "http://www.uniprot.org/taxonomy/9606",

"alternateName":
  {
  "@language": "en",
  "@value": "ABL1_HUMAN"
  },
  {
  "@language": "en",
  "@value": "Proto-oncogene c-Abl"
  }
],
"description": {
  "@language": "en",
  "@value": "Non-receptor tyrosine-protein kinase that plays a role..."
},
"identifier": "http://www.identifiers.org/uniprot/P00519",
"image":
"http://www.identifiers.org/uniprot/P00519#showFeaturesViewer",
```

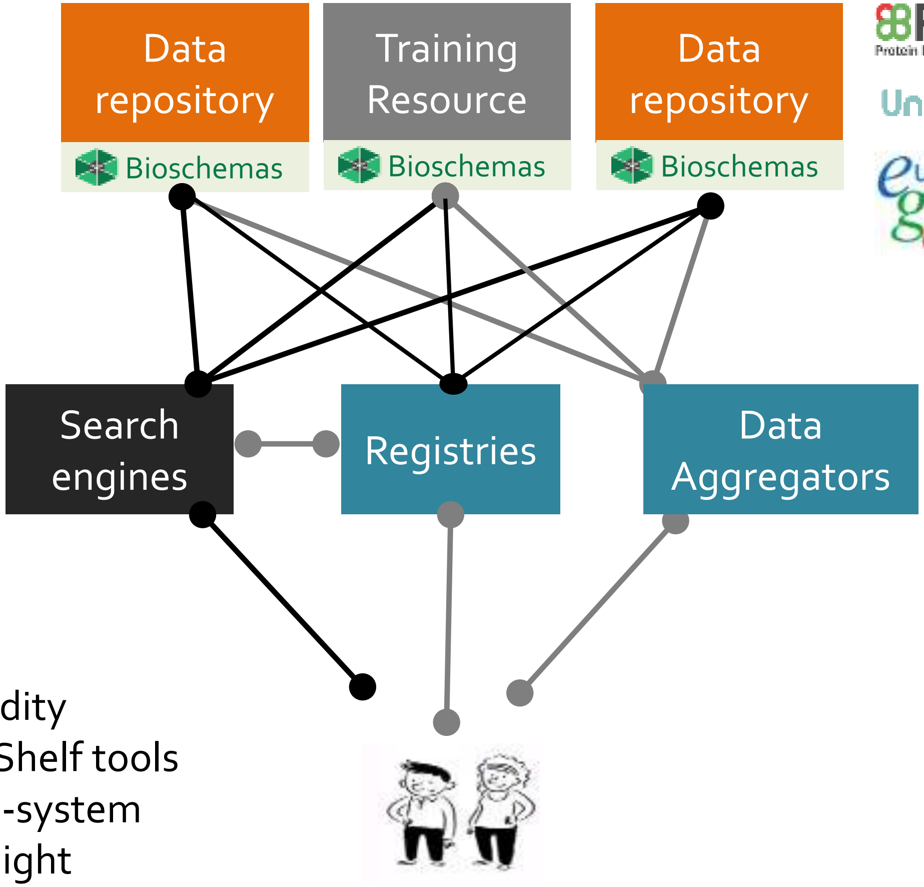
Bioschemas.org

simple ***structured metadata markup*** *on web pages & sitemaps tailored to the Biosciences*

Standardised metadata mark-up

RDFa
JSON-LD
Microdata

Metadata published and harvested without APIs or special feeds



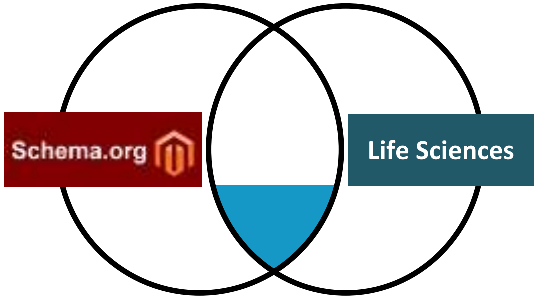
Commodity
Off the Shelf tools
App eco-system
Lightweight

interpro Pfam REACTOME PDBe Protein Data Bank in Europe COPaKB CATH / Gene3D UniProt transPLANT BRASSICA INFORMATION PORTAL European genome-phenome archive THE HUMAN PROTEIN ATLAS

Identifiers.org FAIRsharing.org InterMine OmicsDI

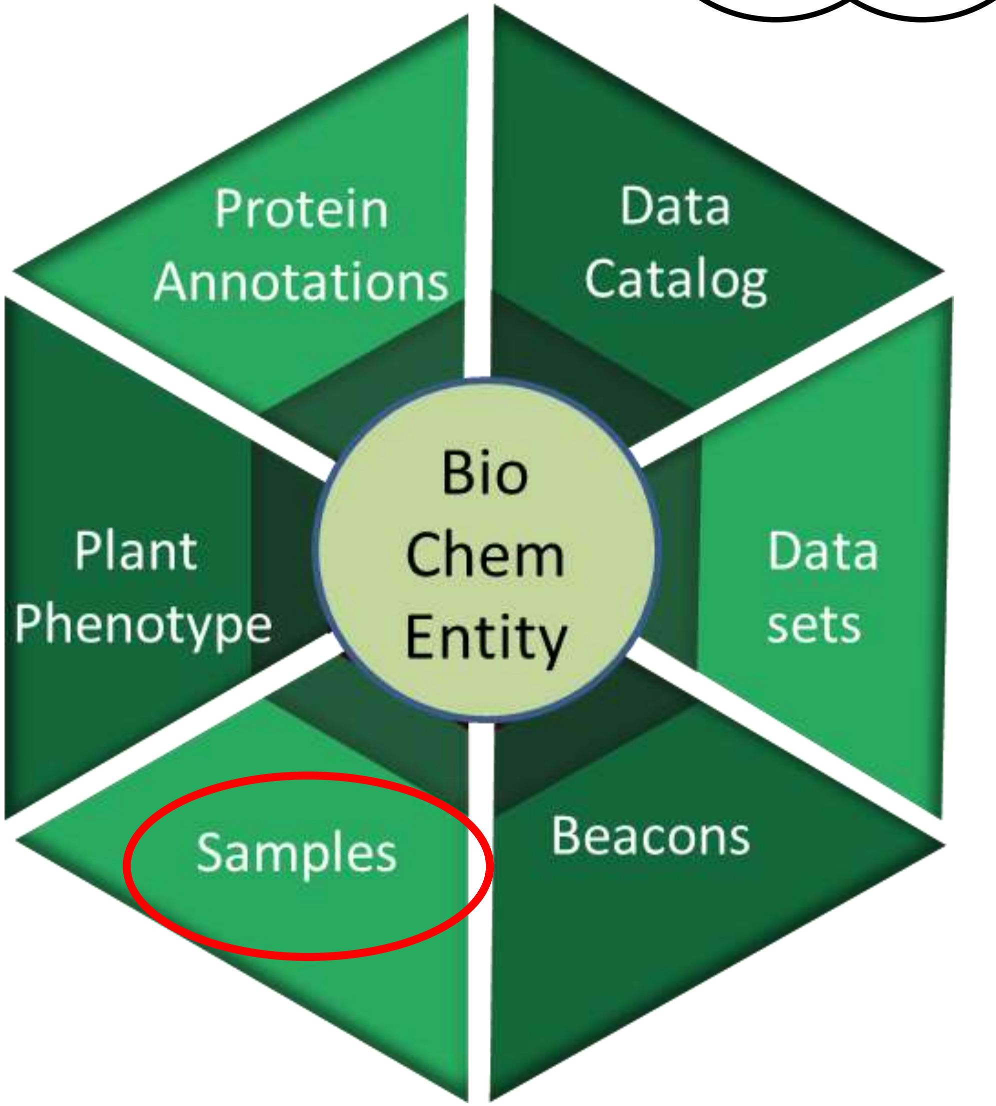
Bioschemas.org

*simple **structured metadata markup** on web pages & sitemaps tailored to the Biosciences*



First specifications:

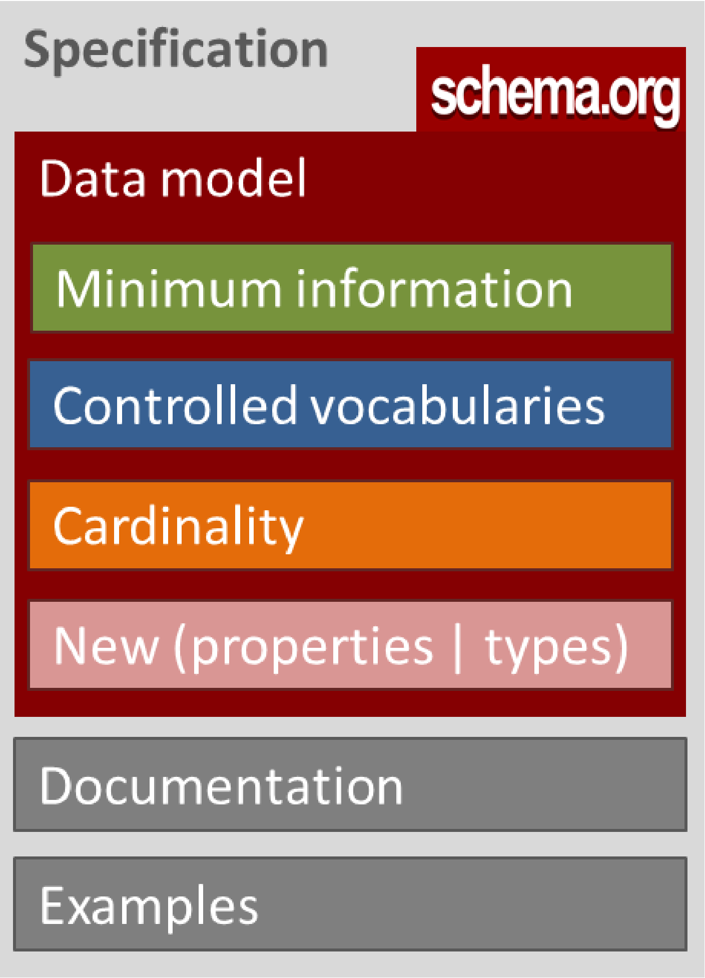
- Bio data infrastructure
  - *DataCatalog*
  - *Datasets*
- Bio data types
  - *Human beacons*
  - ***Samples***
  - *Plant Phenotypes*
  - *Proteins*
  - *(Chemistry)*
- Bio stuff
  - *Training materials*
  - *Events*
  - *Laboratory protocols*
  - *(Workflows and Tools)*



Bioschemas.org

Tailored schema.org to improve
***Findability** and **Accessibility** in Bioscience*

- Specification on top of **schema.org**
- Introduce bioscience types
- Restricted use case
  - Finding data
  - Presenting search results
  - Metadata exchange
- Minimum properties – 6
- Link to domain ontologies not reinventing them



Specification

schema.org

Data model

Minimum information

Controlled vocabularies

Cardinality

New (properties | types)

Documentation

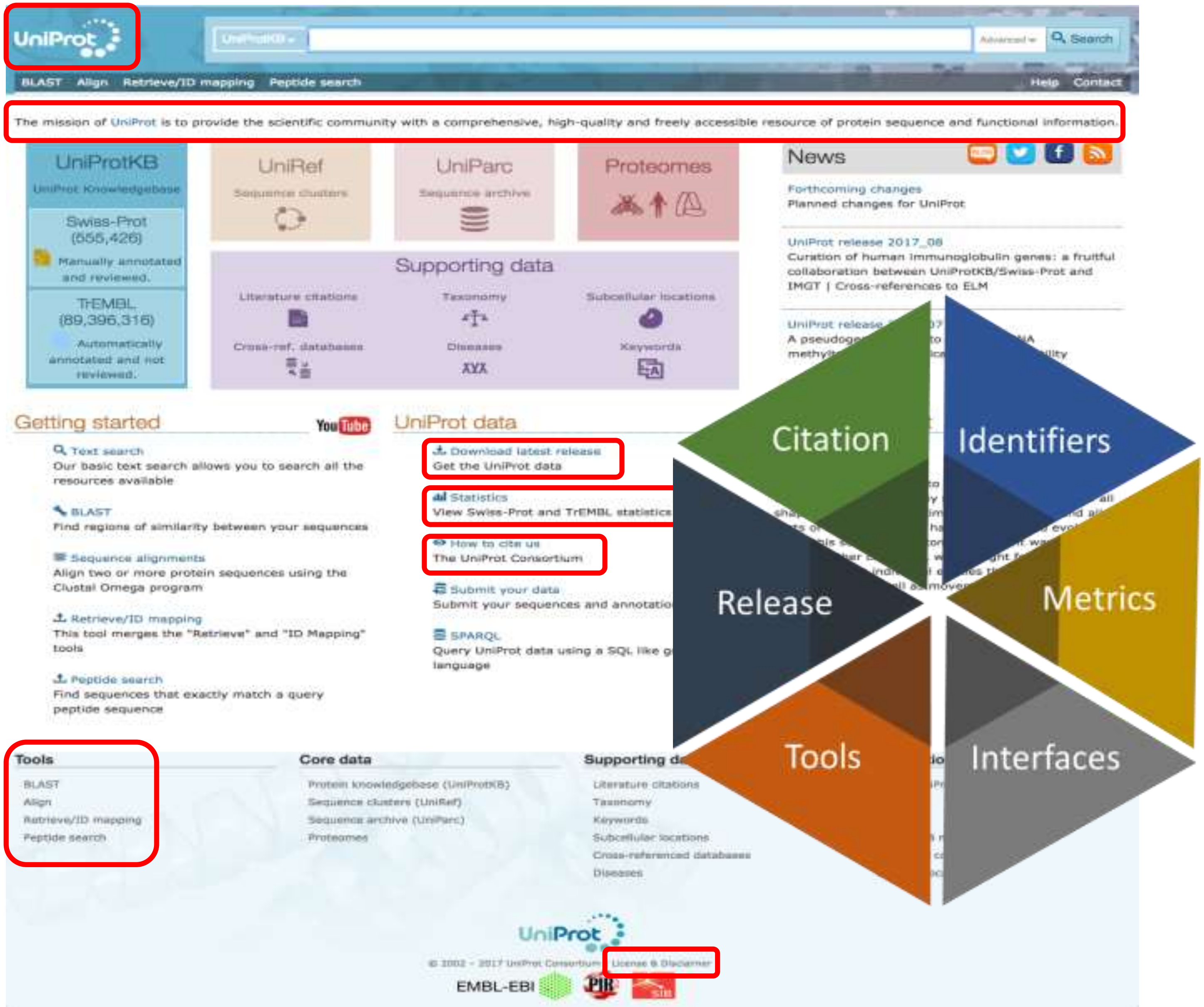
Examples

Layer of constraints +

documentation + extensions

DataCatalog

- Name
- Description
- License
- Release
- Citation
- Metrics
- Tools
- …



UniProt

UniProtKB ▾ | Advanced ▾ | Search

BLAST Align Retrieve/ID mapping Peptide search | Help Contact

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB
UniProt Knowledgebase
Swiss-Prot (555,426)
Manually annotated and reviewed.
TrEMBL (89,396,316)
Automatically annotated and not reviewed.

UniRef
Sequence clusters

UniParc
Sequence archive

Proteomes

Supporting data
Literature citations | Taxonomy | Subcellular locations
Cross-ref. databases | Diseases | Keywords

News

Forthcoming changes
Planned changes for UniProt

UniProt release 2017_08
Curation of human immunoglobulin genes: a fruitful collaboration between UniProtKB/Swiss-Prot and IMGT | Cross-references to ELM

Getting started

Text search
Our basic text search allows you to search all the resources available

BLAST
Find regions of similarity between your sequences

Sequence alignments
Align two or more protein sequences using the Clustal Omega program

Retrieve/ID mapping
This tool merges the "Retrieve" and "ID Mapping" tools

Peptide search
Find sequences that exactly match a query peptide sequence

UniProt data

Download latest release
Get the UniProt data

Statistics
View Swiss-Prot and TrEMBL statistics

How to cite us
The UniProt Consortium

Submit your data
Submit your sequences and annotatio

SPARQL
Query UniProt data using a SQL like g language

Tools
BLAST
Align
Retrieve/ID mapping
Peptide search

Core data
Protein knowledgebase (UniProtKB)
Sequence clusters (UniRef)
Sequence archive (UniParc)
Proteomes

Supporting d
Literature citations
Taxonomy
Keywords
Subcellular locations
Cross-referenced databases
Diseases

UniProt
© 2002 - 2017 UniProt Consortium | License & Disclaimer
EMBL-EBI PIR SIB

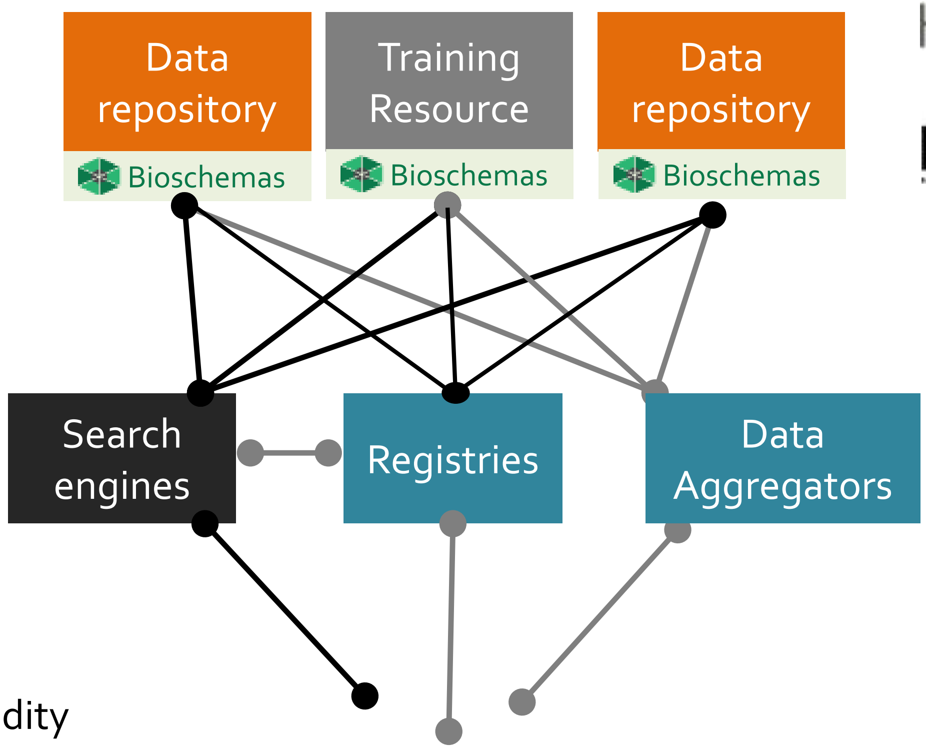
Bioschemas.org

*simple **structured metadata markup** on web pages & sitemaps tailored to the Biosciences*

Standardised metadata mark-up

RDFa
JSON-LD
Microdata

Metadata published and harvested without APIs or special feeds



Commodity
Off the Shelf tools
App eco-system

health RI research infrastructure

BBMRI.nl Biobanking and BioMolecular resources Research Infrastructure

HipSci

FAANG

BIOBANK

BIOBANK

EBiSC

MarRef

BioSamples

Tissue Directory and Coordination Centre

molgenis

RD Connect Sample Catalogue

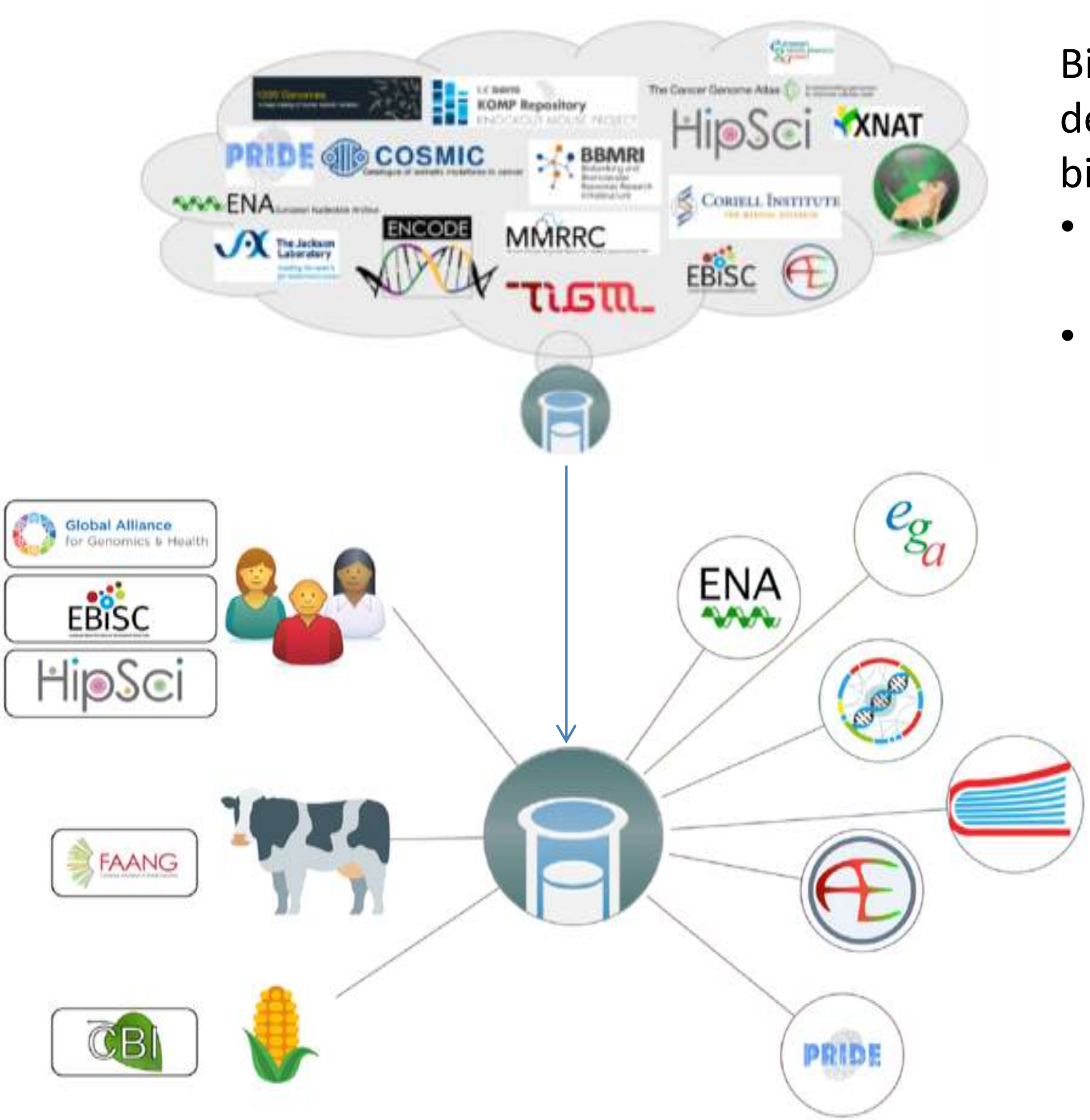
BBMRI-ERIC Directory

Samples

BioSamples at the EBI

BioSamples stores and supplies descriptions and metadata about biological samples.

- 'reference' samples (e.g. from 1000 Genomes, HipSci, FAANG)
- used in an assay database such as the European Nucleotide Archive (ENA) or ArrayExpress.



BioSamples connects across resources

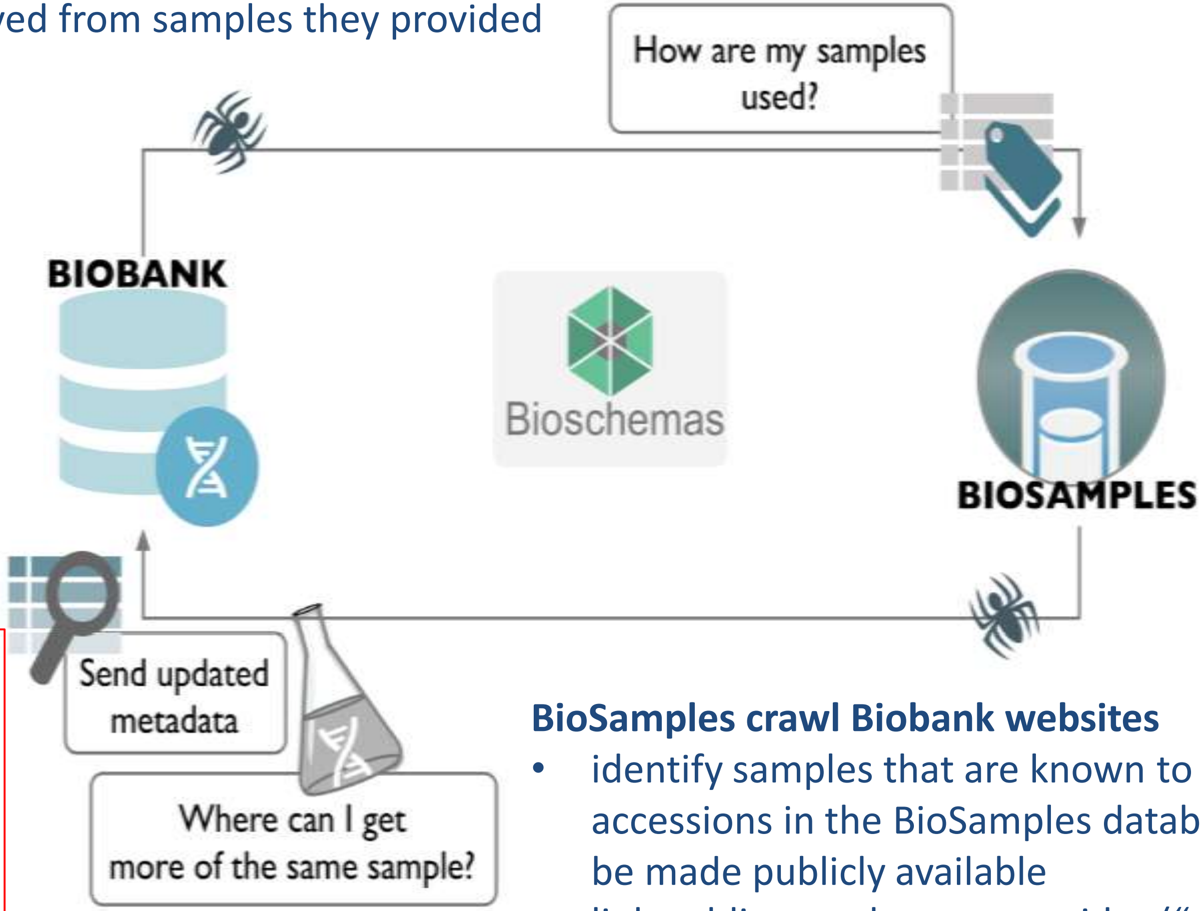
https://www.ebi.ac.uk/biosamples/

Visibility & Credit – Find, Access, and Propagate my biobank metadata

Biobanks crawl BioSamples
identify all the published (and searchable) datasets derived from samples they provided

Biobanks ensure only authorised metadata is visible & control access to restricted samples.

Only Biobanks know the specific samples connected to publicly available datasets



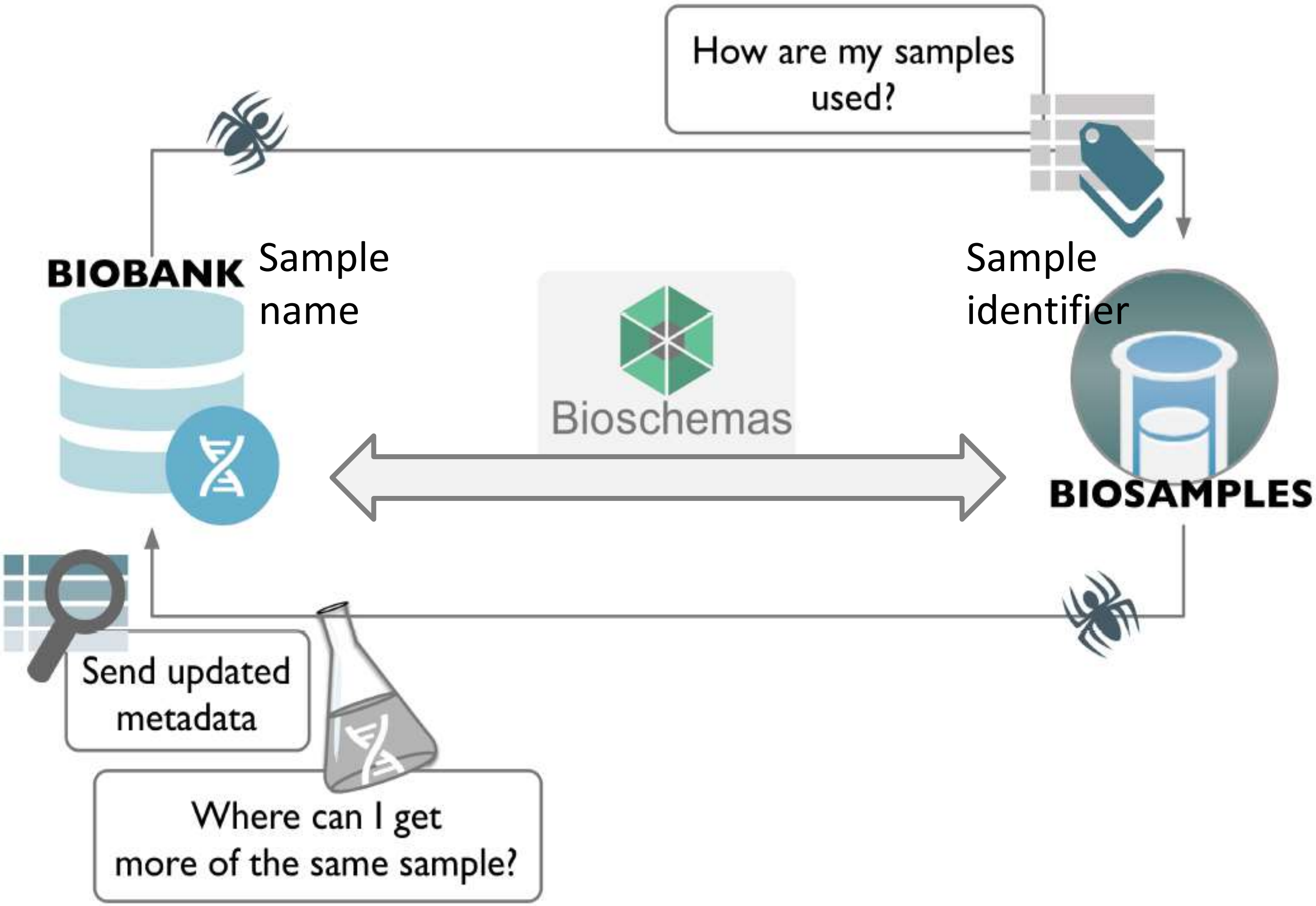
Help ingest sample metadata from data repositories (e.g. Biobank databases) into registries like BioSamples, BBMRI-ERIC Directory, the UKCRC Tissue Directory

BioSamples crawl Biobank websites

- identify samples that are known to have public accessions in the BioSamples database *and* can be made publicly available
- link public samples to a provider (“where can I get more of this sample?”).

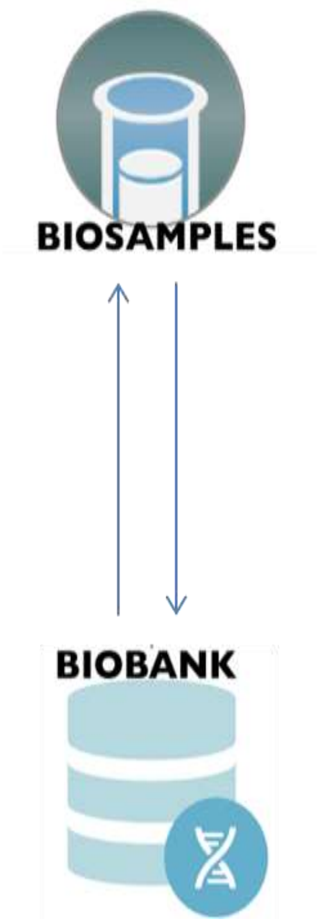
Visibility & Credit

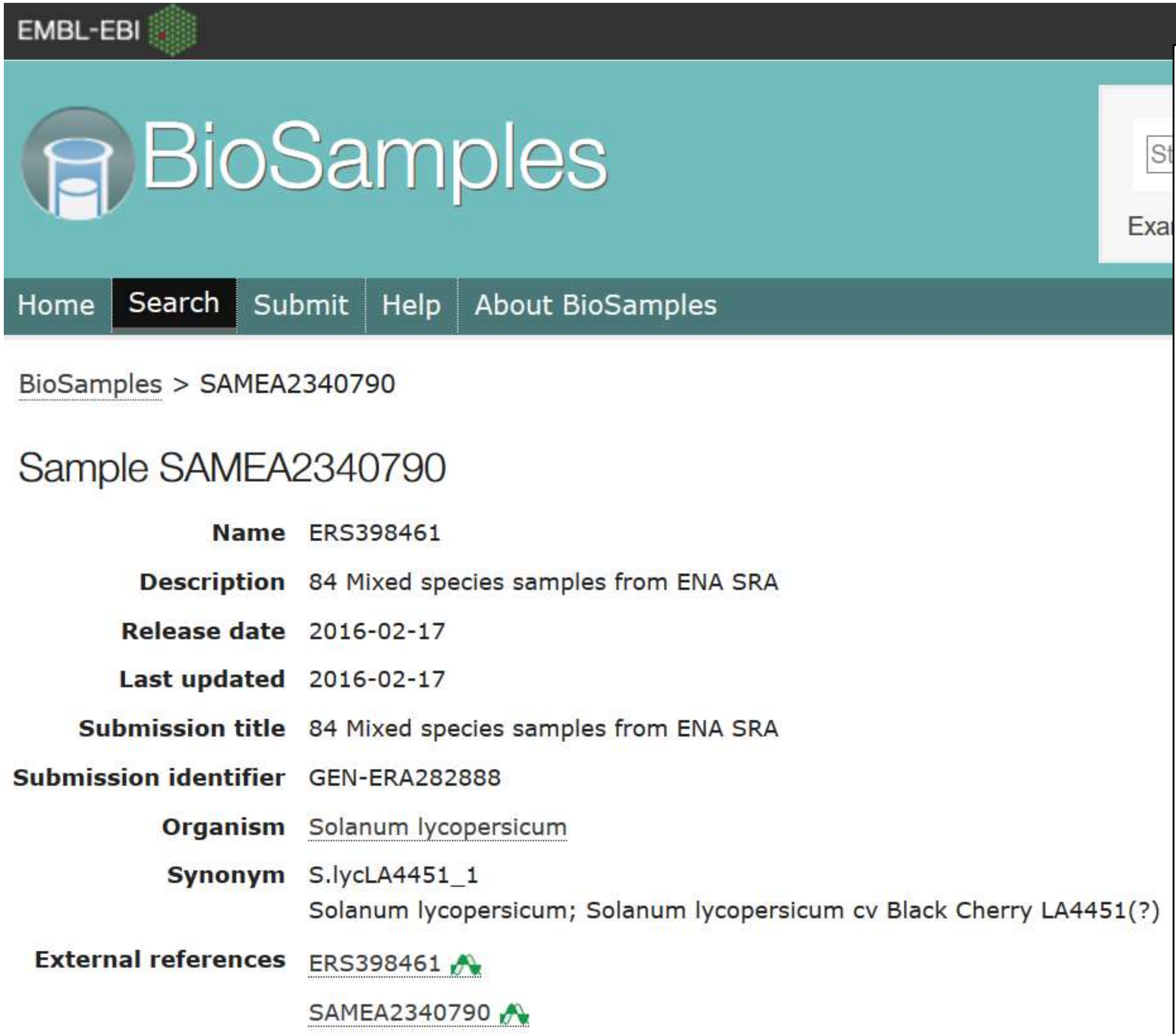
maintain a sanitised bi-directional link through identifiers



Visibility & Credit – Find, Access, and Propagate Biobank metadata

- Customers
  - Direct: Technical providers of biobank and directory platforms
  - Indirect: The users, the Biobanks, the Directories
- Publishing and harvesting
  - Retrospective, prospective and lightweight mechanism easy to incorporate
- Scalable & sustainable
  - Exploits common web infrastructure
  - Standard Search engines can index

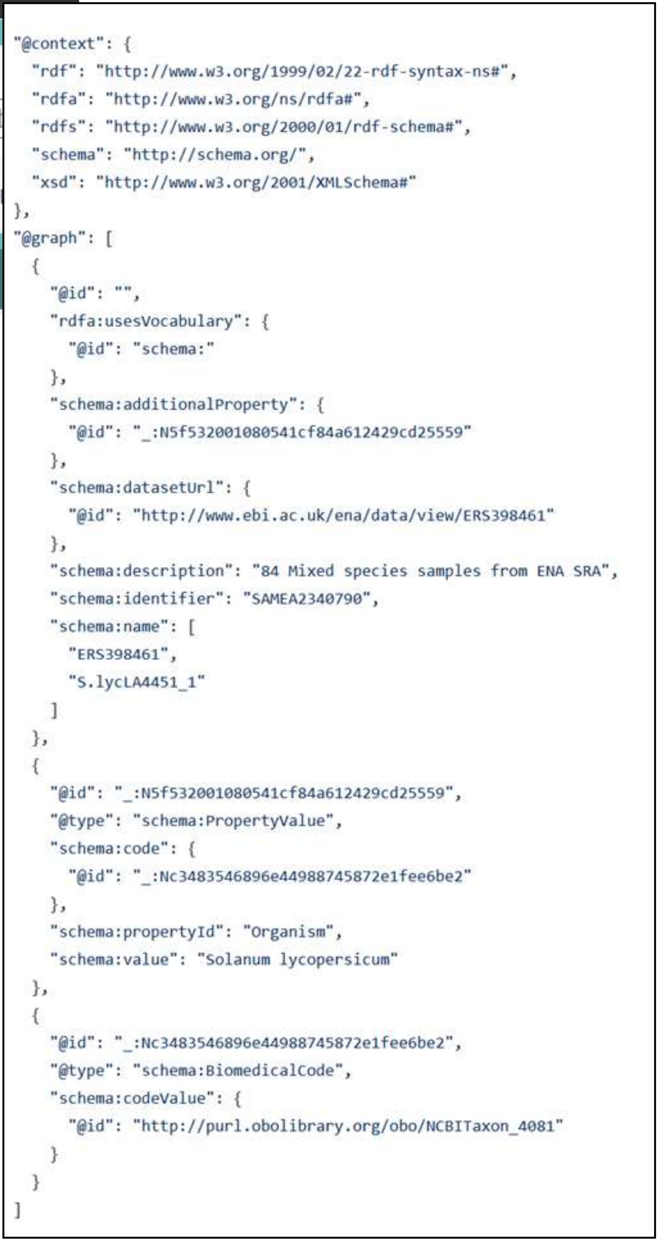




BioSamples > SAMEA2340790

Sample SAMEA2340790

Name	ERS398461
Description	84 Mixed species samples from ENA SRA
Release date	2016-02-17
Last updated	2016-02-17
Submission title	84 Mixed species samples from ENA SRA
Submission identifier	GEN-ERA282888
Organism	Solanum lycopersicum
Synonym	S.lycLA4451_1 Solanum lycopersicum; Solanum lycopersicum cv Black Cherry LA4451(?)
External references	ERS398461 SAMEA2340790



```
"@context": {
  "rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
  "rdfa": "http://www.w3.org/ns/rdfa#",
  "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
  "schema": "http://schema.org/",
  "xsd": "http://www.w3.org/2001/XMLSchema#"
},
"@graph": [
  {
    "@id": "",
    "rdfa:usesVocabulary": {
      "@id": "schema:"
    },
    "schema:additionalProperty": {
      "@id": "_:N5f532001080541cf84a612429cd25559"
    },
    "schema:datasetUrl": {
      "@id": "http://www.ebi.ac.uk/ena/data/view/ERS398461"
    },
    "schema:description": "84 Mixed species samples from ENA SRA",
    "schema:identifier": "SAMEA2340790",
    "schema:name": [
      "ERS398461",
      "S.lycLA4451_1"
    ]
  },
  {
    "@id": "_:N5f532001080541cf84a612429cd25559",
    "@type": "schema:PropertyValue",
    "schema:code": {
      "@id": "_:Nc3483546896e44988745872e1fee6be2"
    },
    "schema:propertyId": "Organism",
    "schema:value": "Solanum lycopersicum"
  },
  {
    "@id": "_:Nc3483546896e44988745872e1fee6be2",
    "@type": "schema:BiomedicalCode",
    "schema:codeValue": {
      "@id": "http://purl.obolibrary.org/obo/NCBITaxon_4081"
    }
  }
]
```

Lightweight, hidden mark-up in the web page

Open Public Process

MIABIS: Minimum Information About BIobank data Sharing (version 2.0)

Find, Cite, Snippets, Metadata exchange

Use cases

schema.org

Mapping

Ideally 6 concepts
Reuse ontologies

Specification

[type] specification [version]

Properties

Examples

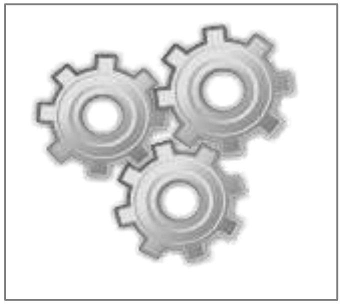
Real mark-up
Tools

Adoption

Protein Data Bank in Europe

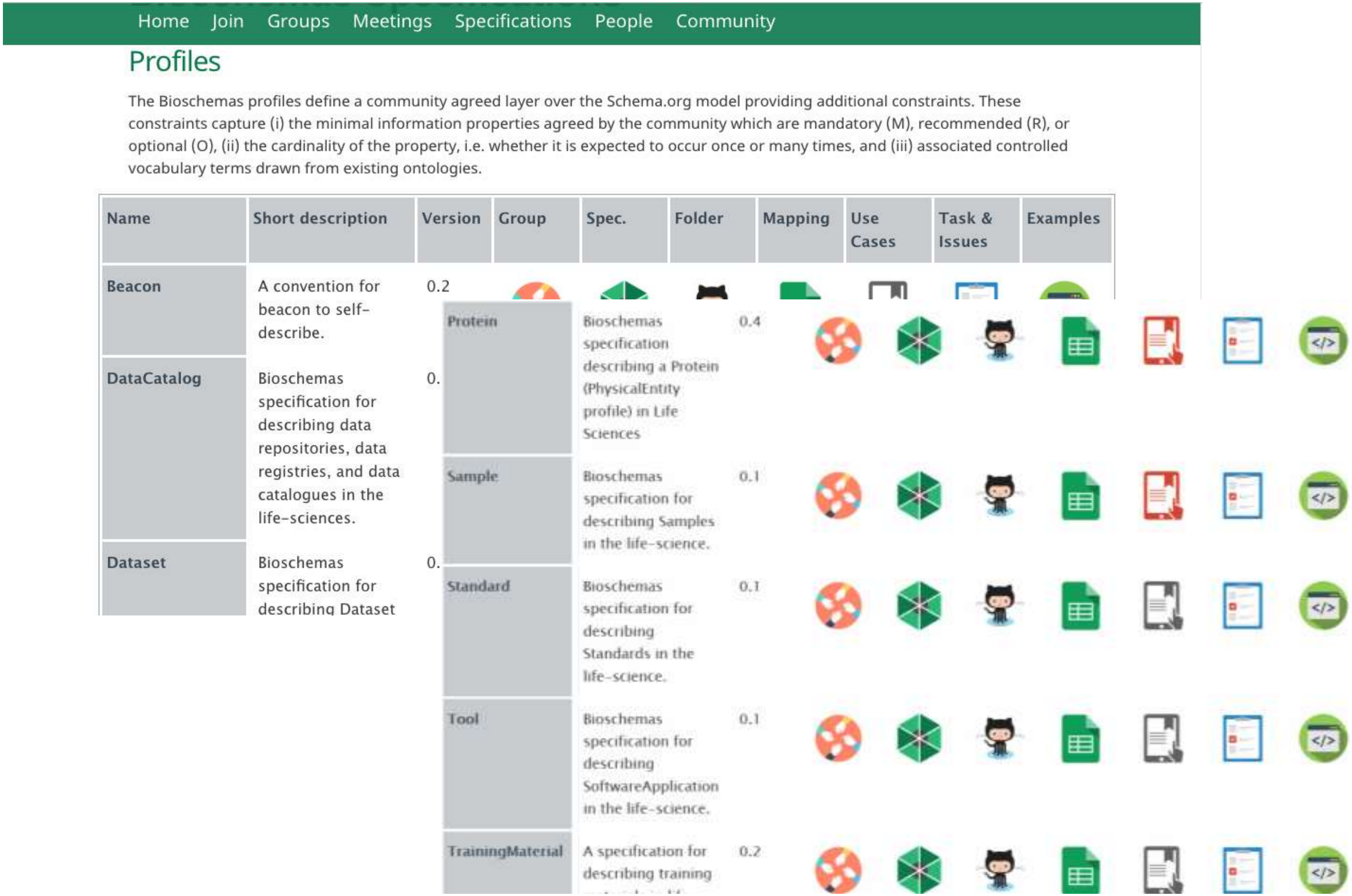
2-3 Oct 2017, Hinxton, ~50 people

Testing



Application

Open Community driven Specifications



Home Join Groups Meetings Specifications People Community

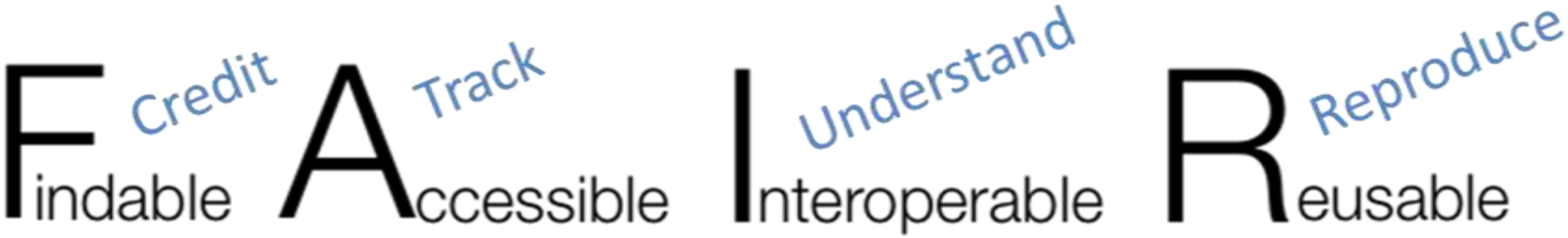
Profiles

The Bioschemas profiles define a community agreed layer over the Schema.org model providing additional constraints. These constraints capture (i) the minimal information properties agreed by the community which are mandatory (M), recommended (R), or optional (O), (ii) the cardinality of the property, i.e. whether it is expected to occur once or many times, and (iii) associated controlled vocabulary terms drawn from existing ontologies.

Name	Short description	Version	Group	Spec.	Folder	Mapping	Use Cases	Task & Issues	Examples
Beacon	A convention for beacon to self-describe.	0.2							
DataCatalog	Bioschemas specification for describing data repositories, data registries, and data catalogues in the life-sciences.	0.							
Dataset	Bioschemas specification for describing Dataset	0.							
Protein	Bioschemas specification describing a Protein (PhysicalEntity profile) in Life Sciences	0.4							
Sample	Bioschemas specification for describing Samples in the life-science.	0.1							
Standard	Bioschemas specification for describing Standards in the life-science.	0.1							
Tool	Bioschemas specification for describing SoftwareApplication in the life-science.	0.1							
TrainingMaterial	A specification for describing training	0.2							

Take Home

Data becomes findable and accessible to the biobank



Create the framework for finding and accessing data through identifiers and metadata exchange using Bioschemas.org

Policies and incentives to get journals, data repositories, funders and authors to mark up and cite.

Acknowledgements

supportive projects and communities

http://elixir-europe.org

http://bioschemas.org

http://fair-dom.org

http://biocaddie.org

http://force11.org

Special thanks: Alasdair Gray, Rafael Jimenez, Tony Burdett, Melanie Courtot, Tim Clark, Julie McMurry, Nick Juty, Sarala Wimalaratne, Norman Morrison, Helen Parkinson, Leyla Garcia, Susanna Sansone, Philippe Rocca-Serra, Alejandra Gonzalez-Beltran, Phil Quinlan