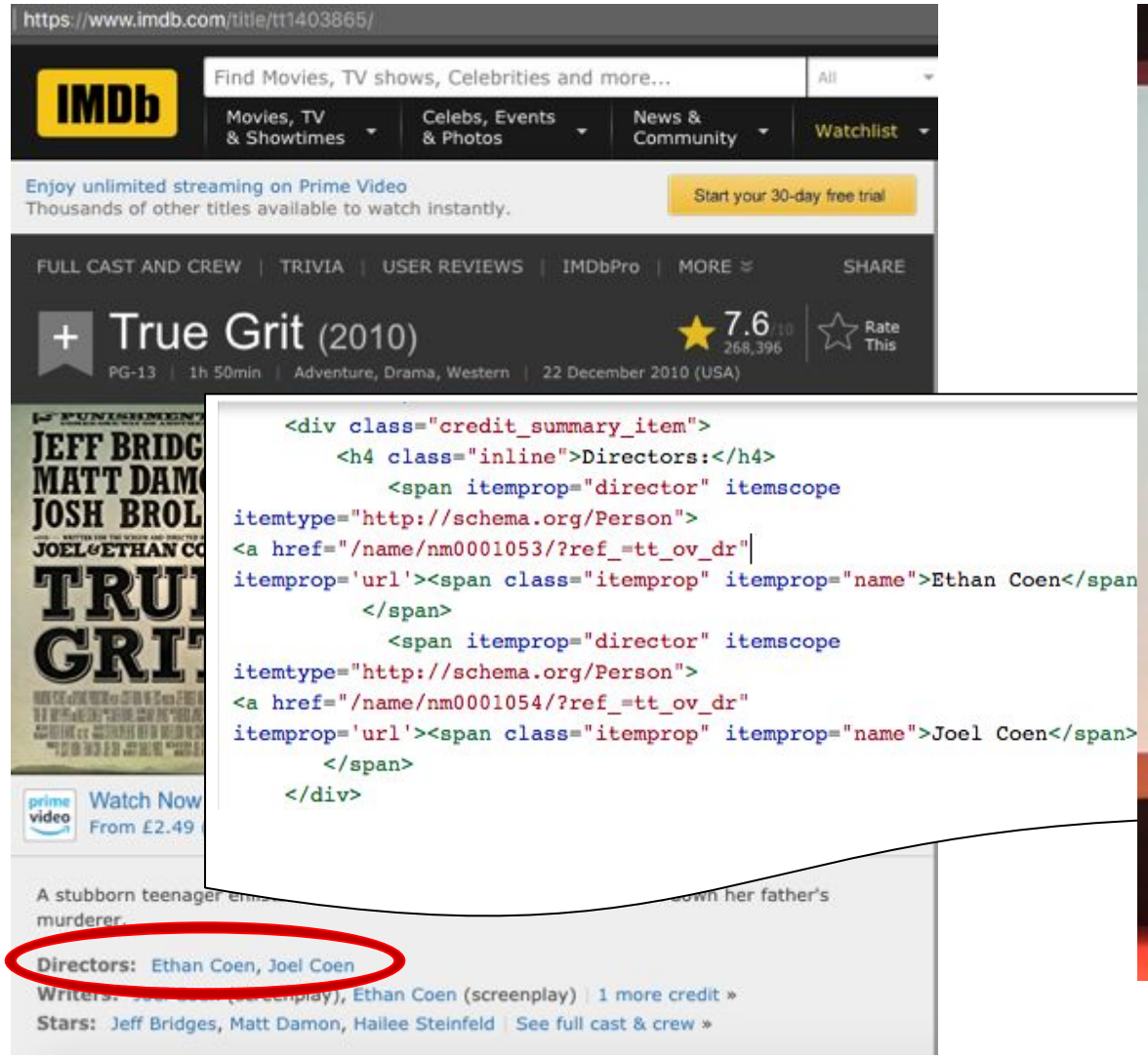
Discovering Biodiversity Resources on the Web

Alasdair Gray, Franck Michel, Quentin Groom, and the Bioschemas Community

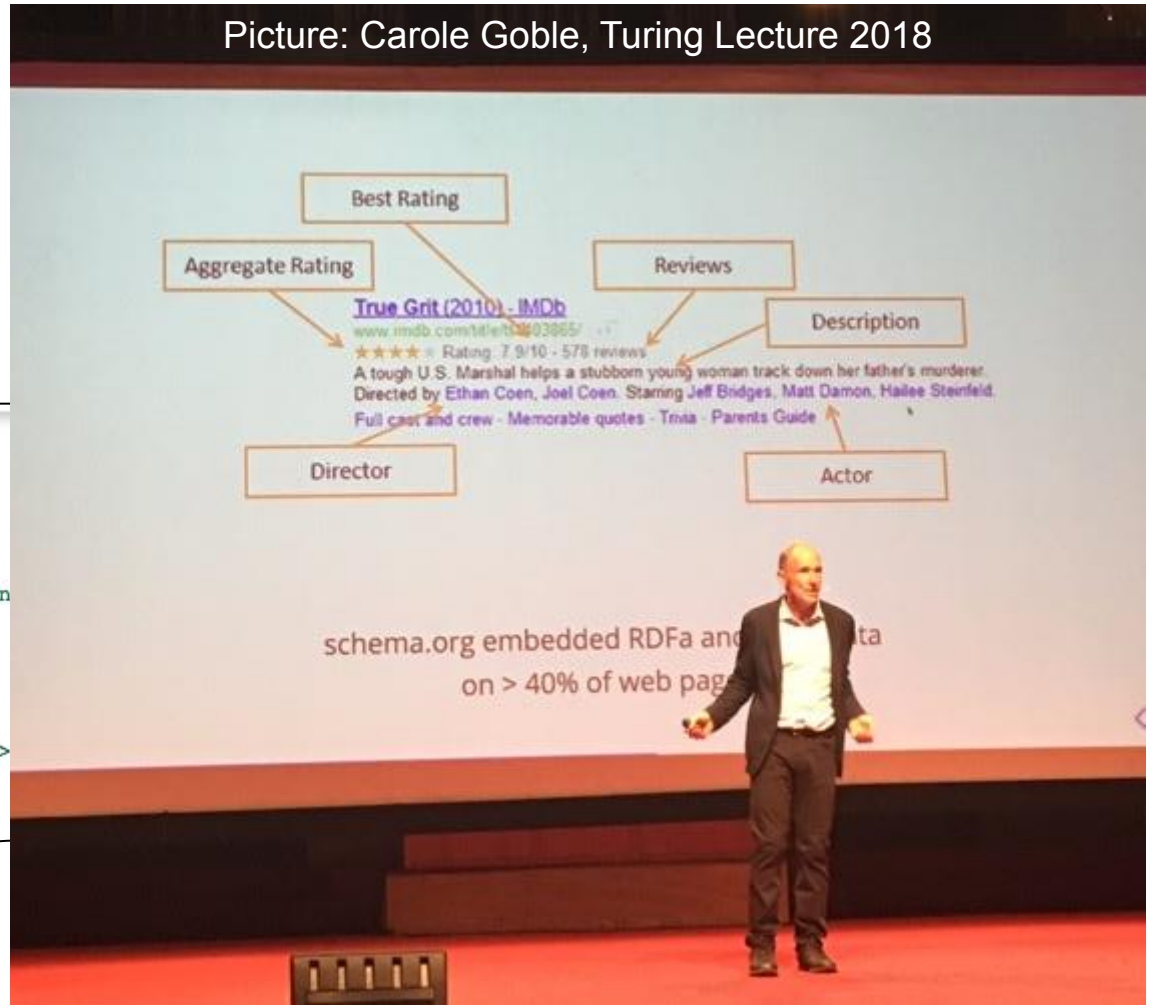
Alasdair J.G. Gray
Bioschemas Steering Council Chair
Heriot-Watt University – ELIXIR-UK

Biodiversity Mini-symposium
ELIXIR All Hands, 8 June 2020

Schema.org: Semantic Markup for the Web




```
    <div class="credit_summary_item">
        <h4 class="inline">Directors:</h4>
            <span itemprop="director" itemscope
itemtype="http://schema.org/Person">
<a href="/name/nm0001053/?ref_=tt_ov_dr"
itemprop='url'><span class="itemprop" itemprop="name">Ethan Coen</span
          </span>
            <span itemprop="director" itemscope
itemtype="http://schema.org/Person">
<a href="/name/nm0001054/?ref_=tt_ov_dr"
itemprop='url'><span class="itemprop" itemprop="name">Joel Coen</span>
       </span>
    </div>
```



Structured data → descriptors

What we are talking about

Types (829)

- Thing
  - Action
    - AchieveAction
      - LoseAction
      - TieAction
      - WinAction
    - AssessAction
      - ChooseAction
        - VoteAction
      - IgnoreAction
      - ReactAction
        - AgreeAction
        - DisagreeAction
        - DislikeAction
        - EndorseAction
        - LikeAction
        - WantAction
      - ReviewAction
    - ConsumeAction
      - DrinkAction
      - EatAction

What we can say about those things

Properties (1351)

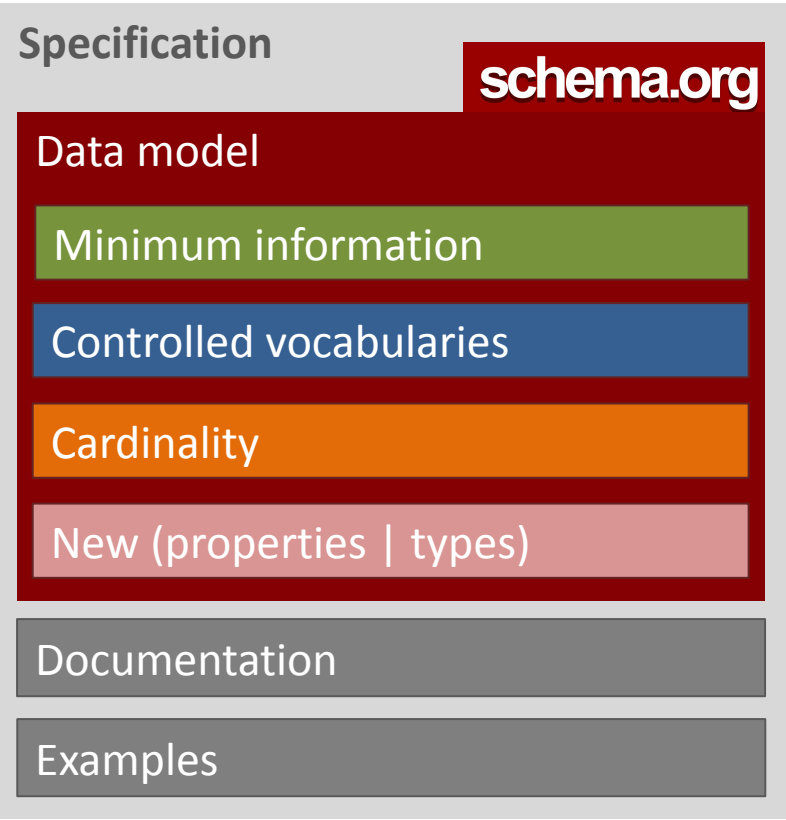
Property	Expected Type	Description
Properties from Person		
additionalName	Text	An additional name for a Person, can be used for a middle name.
address	PostalAddress or Text	Physical address of the item.
affiliation	Organization	An organization that this person is affiliated with. For example, a school/university, a club, or a team.
alumniOf	EducationalOrganization or Organization	An organization that the person is an alumni of. Inverse property: alumni.
award	Text	An award won by or for this item. Supersedes awards.
birthDate	Date	Date of birth.
birthPlace	Place	The place where the person was born.
brand	Brand or Organization	The brand(s) associated with a product or service, or the brand(s) maintained by an organization or business person.
children	Person	A child of the person.
colleague	Person or URL	A colleague of the person. Supersedes colleagues.
contactPoint	ContactPoint	A contact point for a person or organization. Supersedes contactPoints.
deathDate	Date	Date of death.
deathPlace	Place	The place where the person died.
duns	Text	The Dun & Bradstreet DUNS number for identifying an organization or business person.
email	Text	Email address.
familyName	Text	Family name. In the U.S., the last name of an Person. This can be used along with givenName instead of the name property.
faxNumber	Text	The fax number.
follows	Person	The most generic uni-directional social relation.
	Organization or	A person or organization that supports (sponsors) something

Bioschemas

- Community initiative built on top of **schema.org**
- Aim
  - Improve data discoverability and interoperability in Life Sciences
- Approach
  - Add Life Science types to schema.org
  - Provide usage guidelines and examples
    - 6 Minimal properties
    - Link to domain ontologies
  - Support software

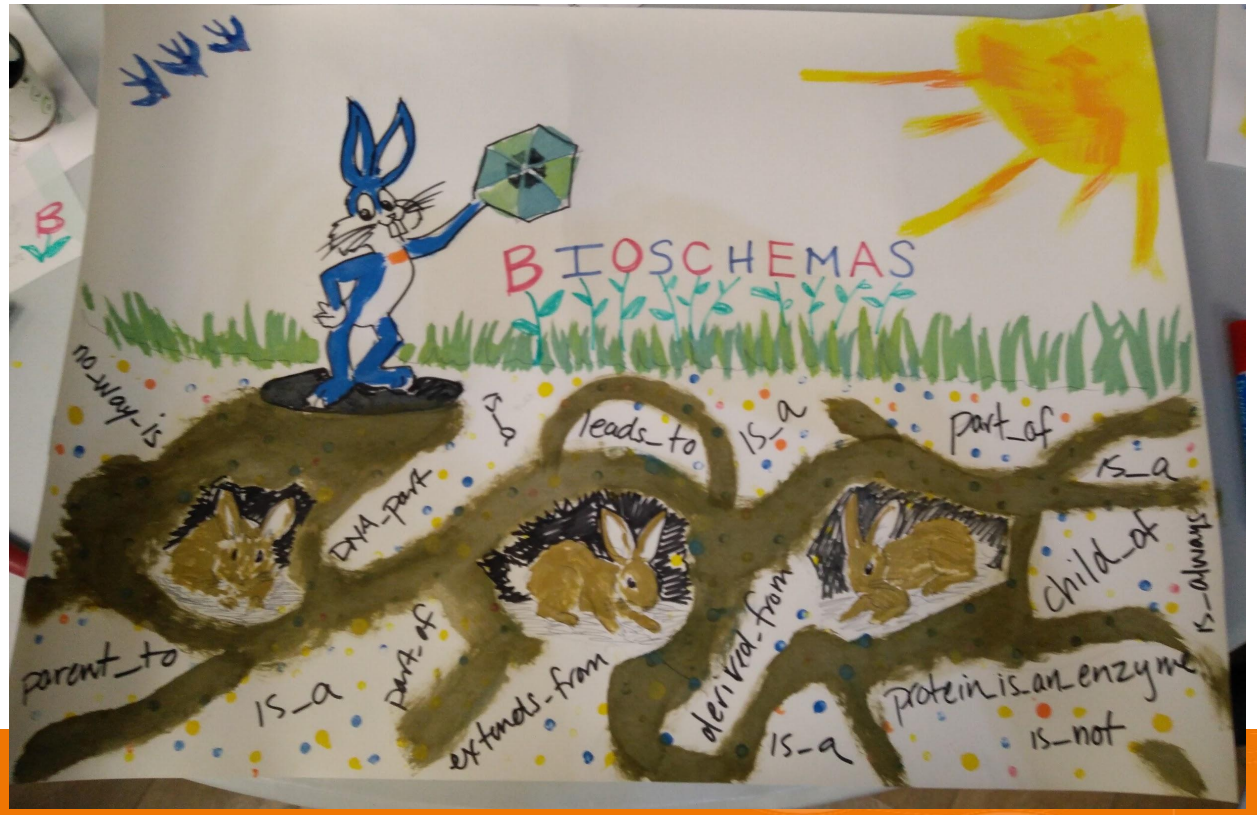
Profile over **schema.org**

Layer of constraints + documentation + extensions



Bioschemas: Lightweight semantic interoperability

- Many domain ontologies
  - Designed to model biology
- Bioschemas focus on search!



BioPortal

BioPortal Statistics	
Ontologies	849
Classes	11,336,085
Resources Indexed	48
Indexed Records	39,537,360
Direct Annotations	95,468,433,792
Direct Plus Expanded Annotations	144,789,582,932

OLS

Data Content

Updated 08 May 2020 04:38

- 251 ontologies
- 6,198,216 terms
- 27,921 properties
- 486,139 individuals

Findable

- ★Globally unique identifiers
- ★Community defined enriched metadata
- ★Indexable by search engines

Accessible

- ★Retrievable
- ★HTTP

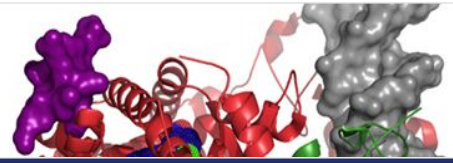
Interoperable

- ★JSON-LD/RDFa
- ★Link to controlled vocabularies
- ★Links to other resources

Reusable

- ★ License
- ★ Provenance

Bioschemas: Markup for the Life Sciences



IUPHAR/BPS
Guide to PHARMACOLOGY

Search Database

Home About Help

Search GtoPdb...

Ligands relevant to SARS-CoV-2(COVID-19) - ligand name links to detailed information in GtoPdb, or to our pre-release blog.

Download as CSV

Ligand (Therapeutic)	Ligand ID	Comments
Abl kinase inhibitors (e.g. imatinib)	5687	Abelson kinase (Abl) inhibitors are reported to block Spike protein-induced SARS-CoV and MERS-
acalabruntinib		
ACE2 ligands		
anakinra		
anti-TNF therapy (e.g. inflixi nd others)		
apilimod		
ASC09F		

```
<!--- BioSchemas ---->
<script type="application/ld+json">[

{
  "@context": "https://schema.org",
  "@type": "Protein",
  "@id": "https://www.guidetopharmacology.org/GRAC/ObjectDisplayForward?objectId=16
  "http://purl.org/dc/terms/conformsTo": "https://bioschemas.org/profiles/Protein/
  "identifier": "1614",
  "name": "ACE2",
  "associatedDisease": "COVID-19",
  "description": "Receptor on host cells that is exploited by some betacoronaviruses for viral entry. Engaged by
  SARS-CoV-2 spike protein as the first step towards infection of host cells",
  "url": "https://www.guidetopharmacology.org/GRAC/ObjectDisplayForward?objectId=1614"
},
{
  "@context": "https://schema.org",
  "@type": "MolecularEntity",
  "@id": "https://www.guidetopharmacology.org/GRAC/LigandDisplayForward?ligandId=8912#",
  "http://purl.org/dc/terms/conformsTo": "https://bioschemas.org/profiles/MolecularEntity/0.4-DRAFT-2019_11_11/",
  "identifier": "8912",
  "name": "acalabruntinib",
  "url": "https://www.guidetopharmacology.org/GRAC/LigandDisplayForward?ligandId=8912",
  "associatedDisease": "COVID-19",
  "description": "acalabrutinib is an approved BTK inhibitor that is
  progressed to clinical trial in COVID-19 patients as part
```

FAIR

DPC

FAIDARE

orphanet

The DataCite
PID Graph

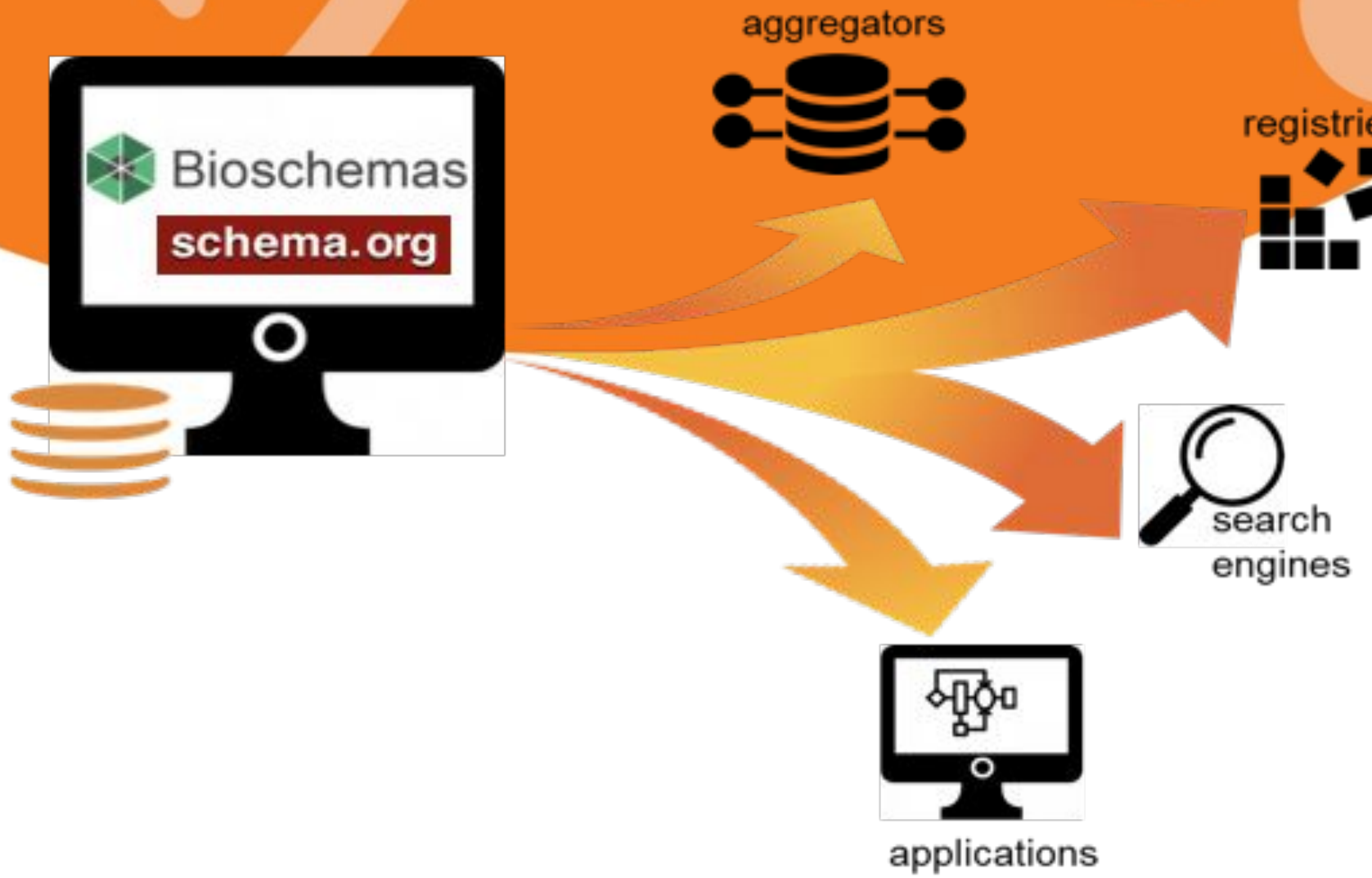
OPEN
RESEARCH KNOWLEDGE GRAPH
DIGITAL LIBRARIES FOR SEMANTIC SCIENTIFIC KNOWLEDGE

OpenAIRE
Research Graph

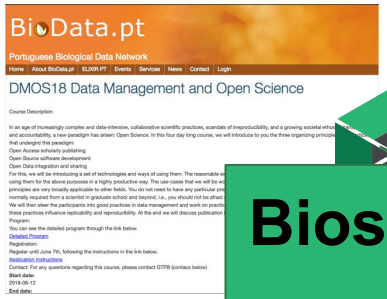
ResearchGraph
http://researchgraph.org/

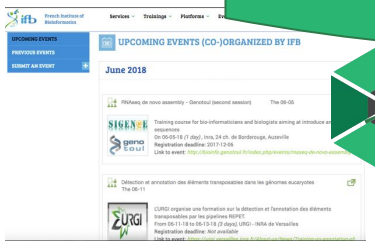
elixir

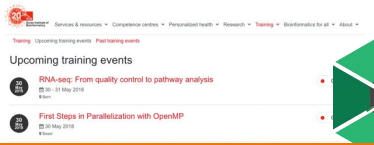
Exploiting Bioschemas Markup



TeSS: Specialised Search

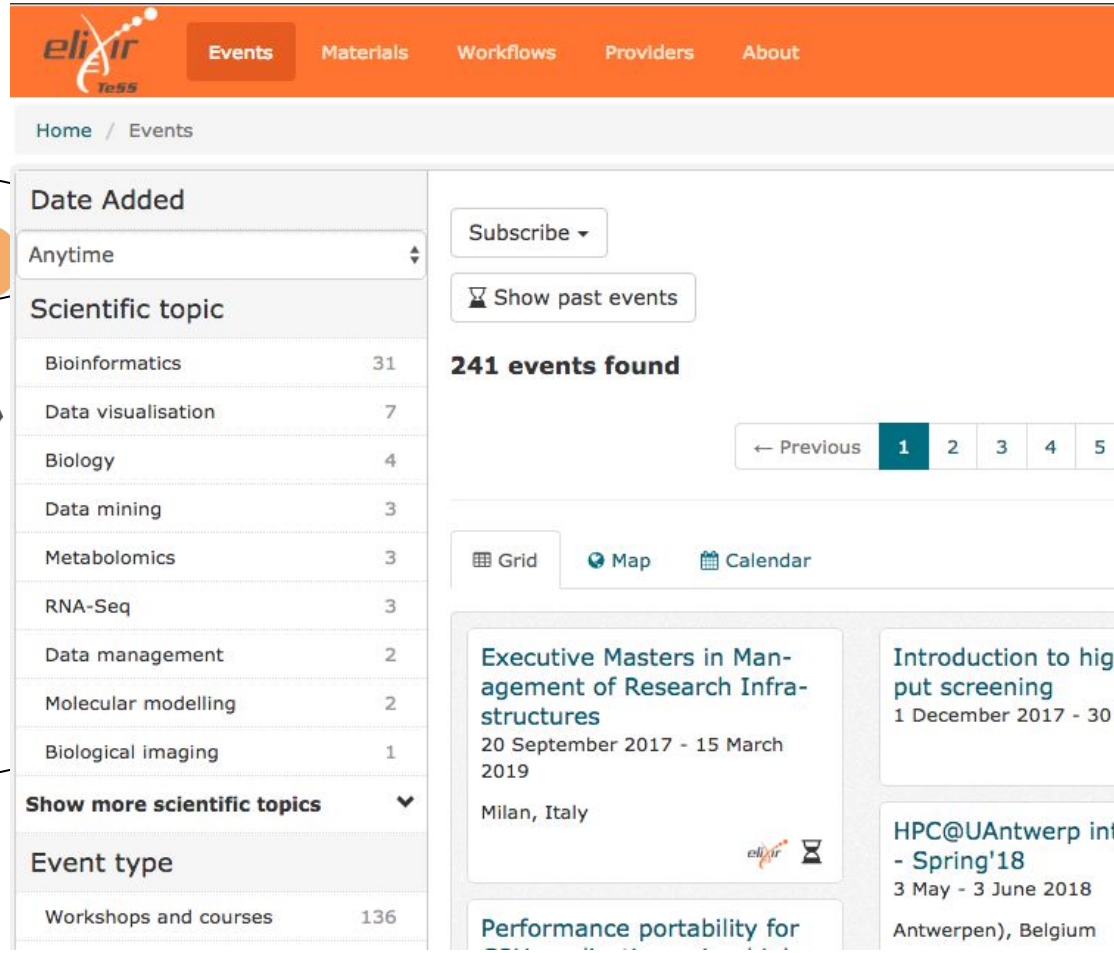






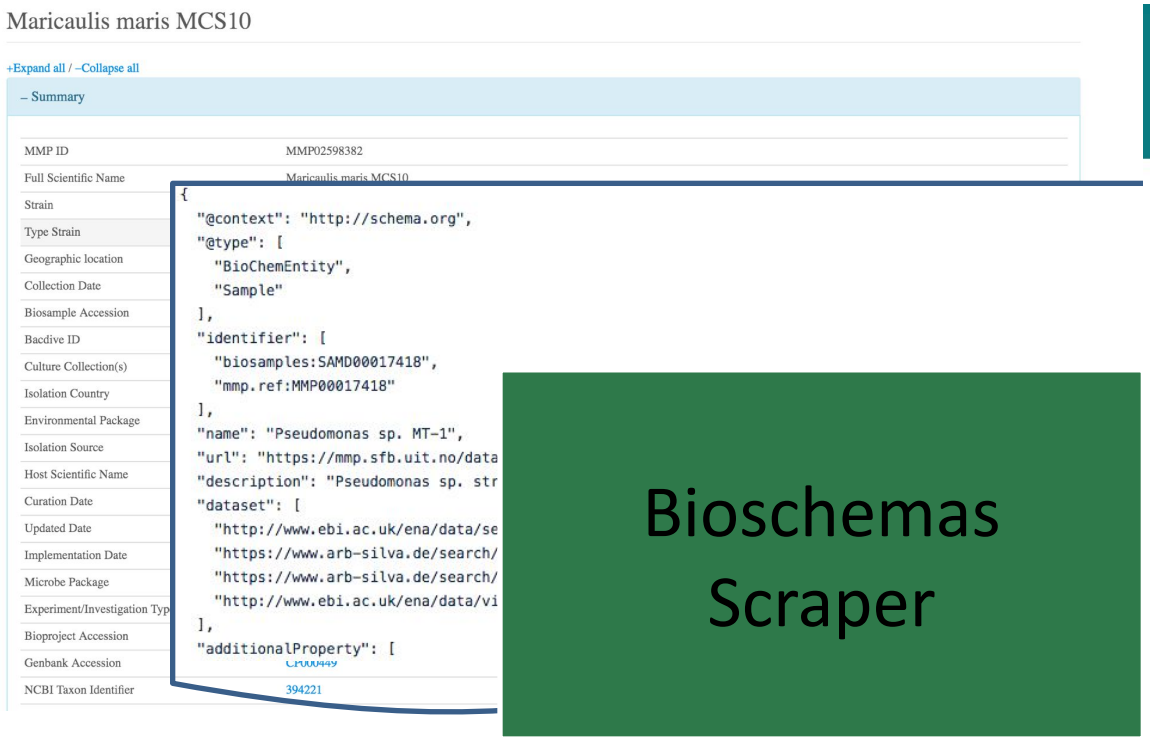
Bioschemas Event:

- contact
- description
- endDate
- eventType
- hostInstitution
- location
- name
- startDate
- …

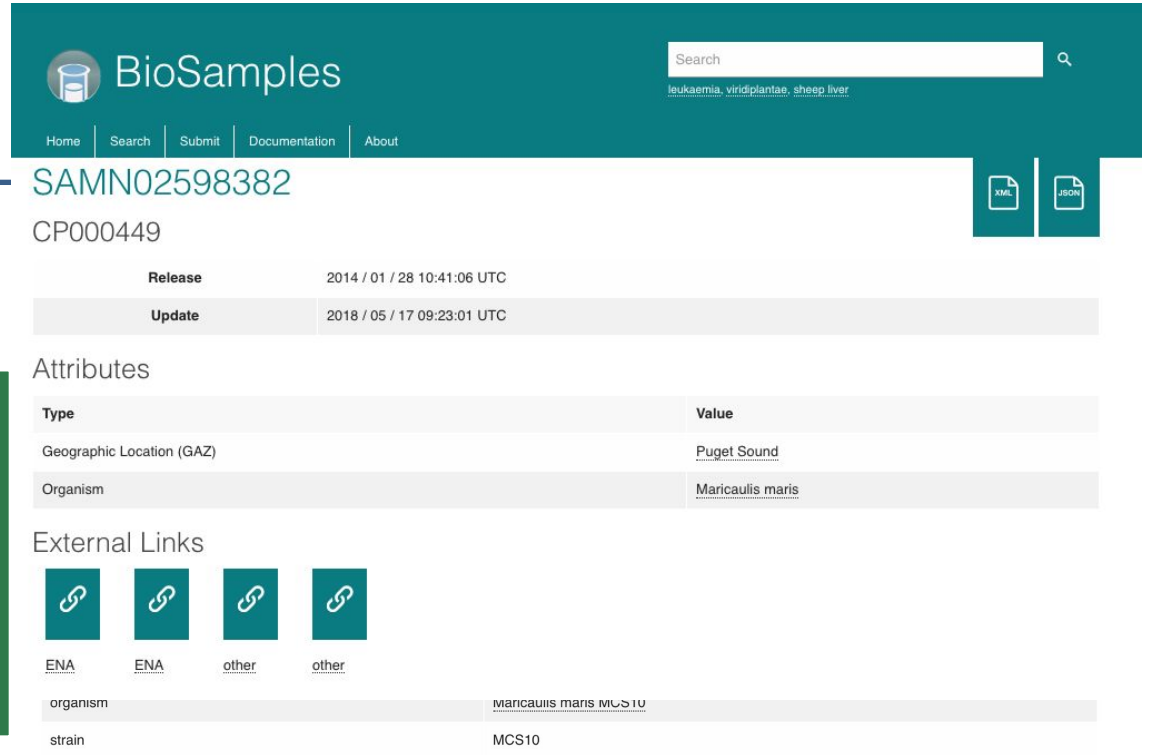


Data Exchange: Without an API

MarRef → BioSamples

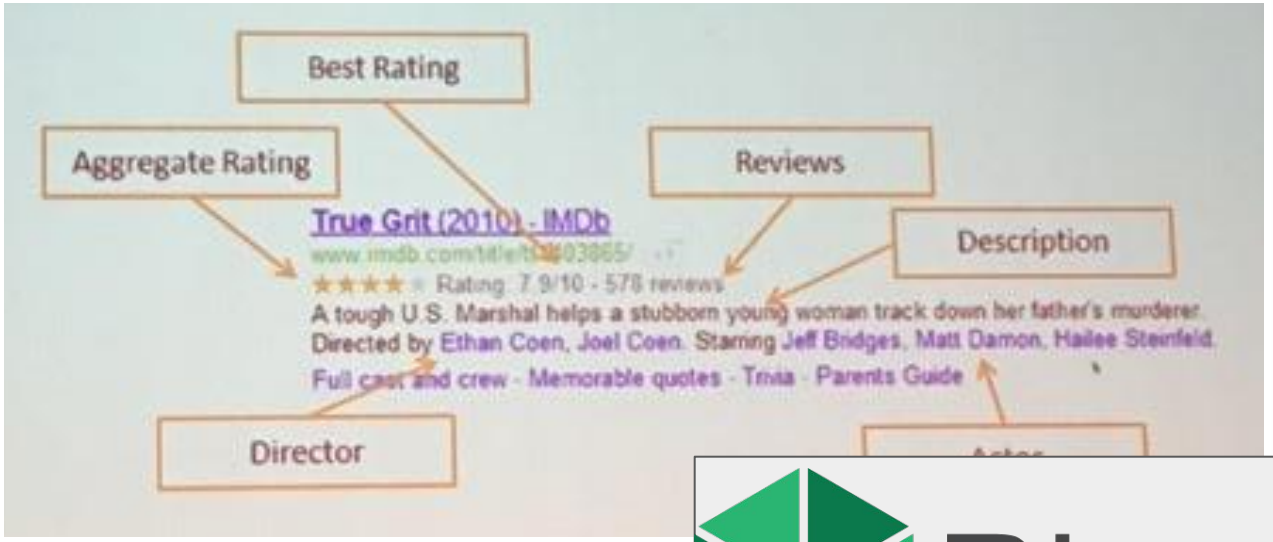


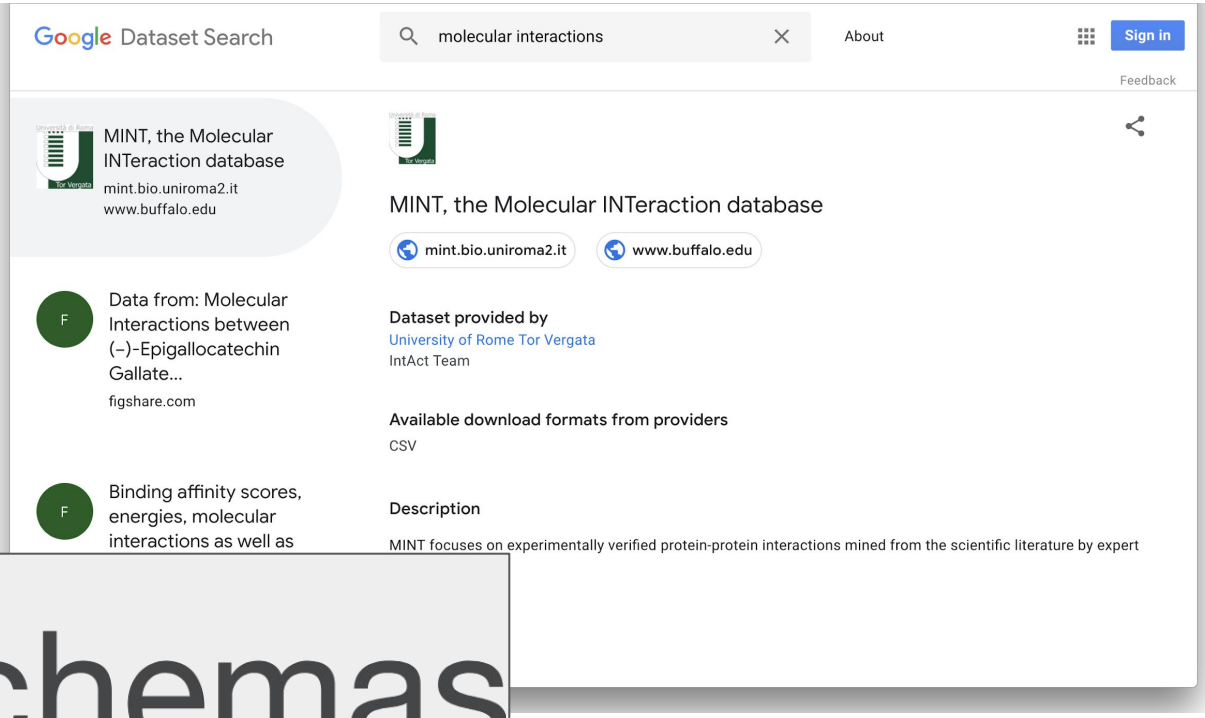
Bioschemas Scraper



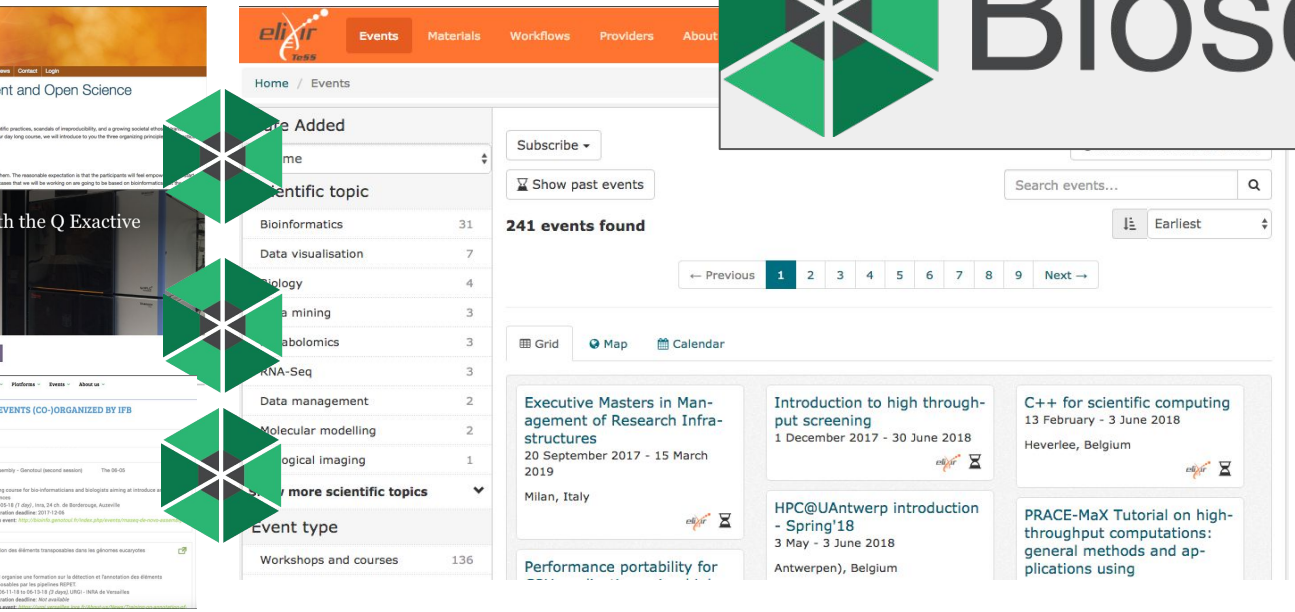
https://github.com/EBIBioSamples/bioschemas_marref_demo/blob/master/Summary.md

Rich snippet generation





Bioschemas



The DataCite PID Graph

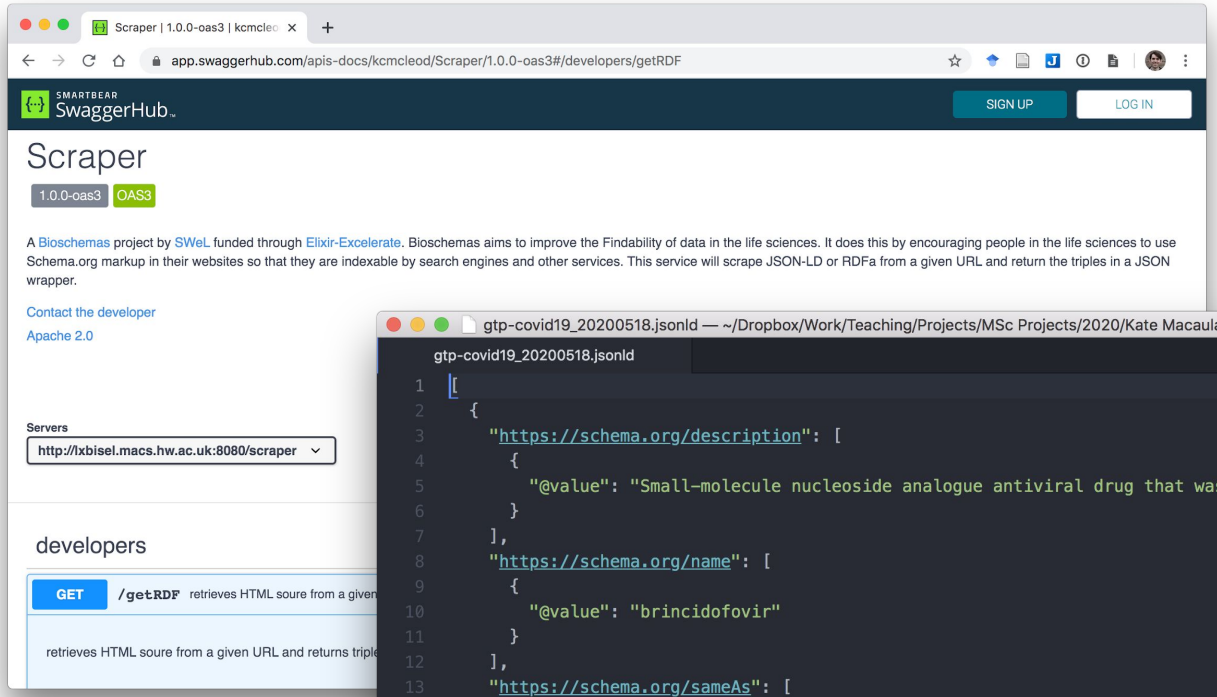
OPEN RESEARCH KNOWLEDGE GRAPH
DIGITAL LIBRARIES FOR SEMANTIC SCIENTIFIC KNOWLEDGE

OpenAIRE Research Graph

ResearchGraph
http://researchgraph.org/

BMUSE: Bioschemas Markup Scraper and Extractor

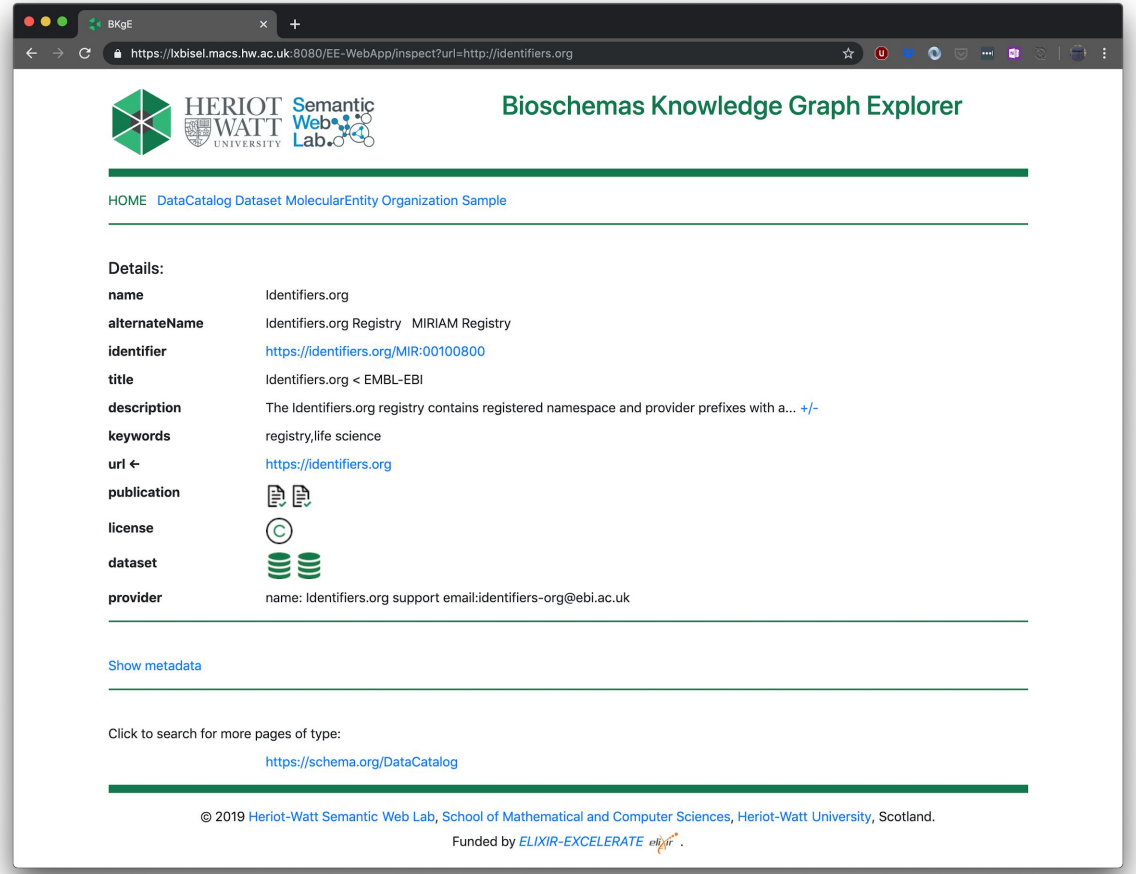
- Directed scraper
  - List of URLs
  - Site maps
- Extracts markup
  - JSON-LD or RDFa
  - Single page applications
- Includes provenance of extraction
  - retrievedFrom
  - retrievedOn
- Experimental crawl of COVID-19 markup

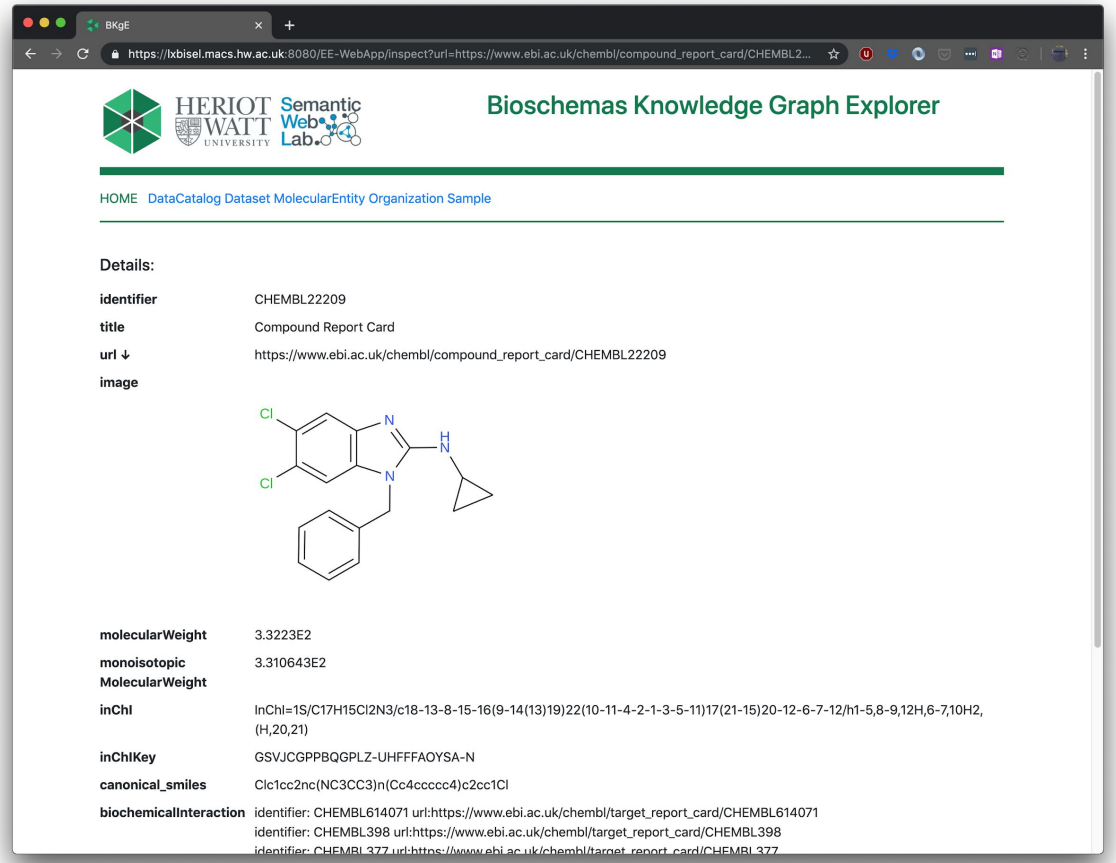


https://github.com/HW-SWeL/BMUSE

BKG Explorer

Built over Bioschemas markup crawled from 30 live deployments
20,000 pages





Bioschemas for Biodiversity

aggregators

registries

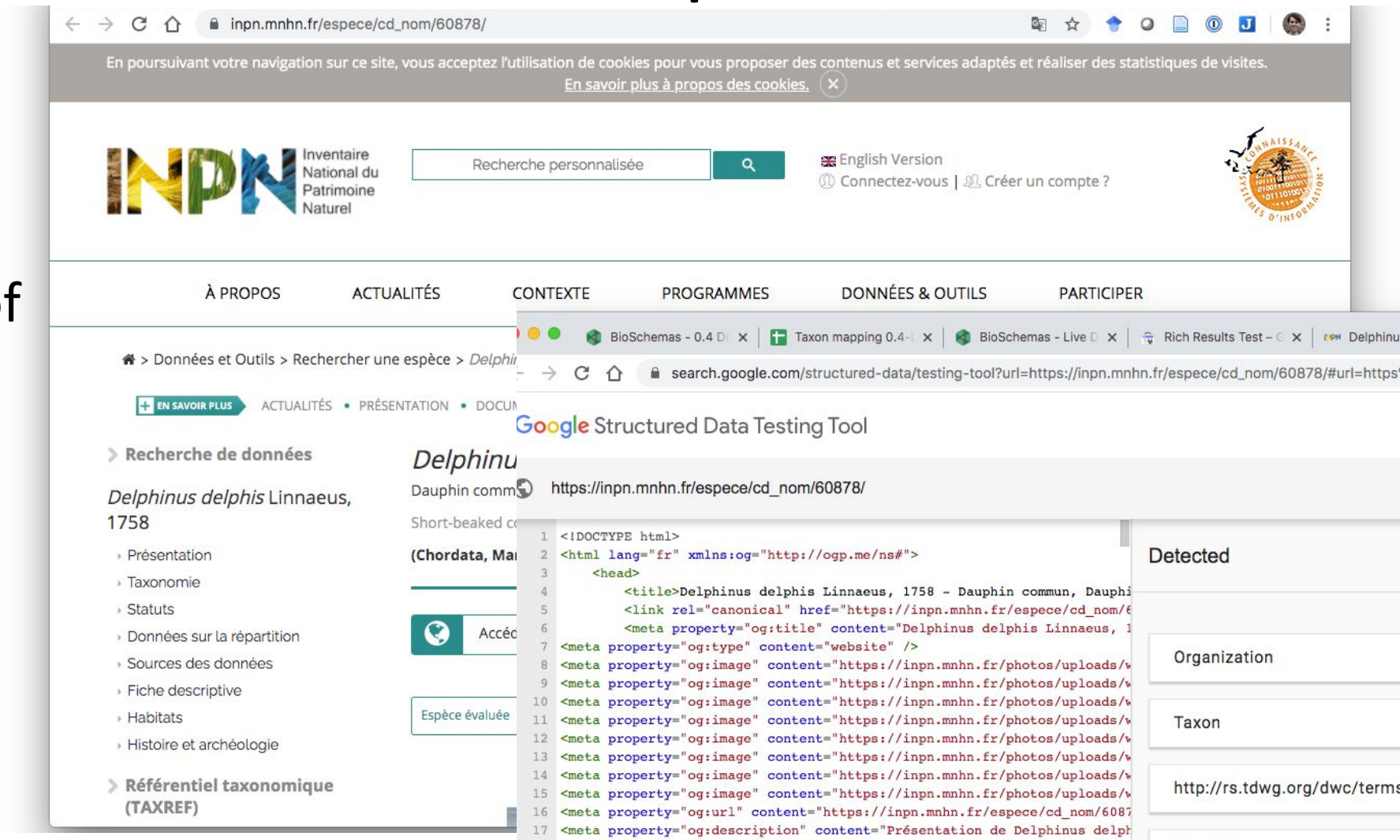
search engines

applications

Image: https://en.wikipedia.org/wiki/Biodiversity#/media/File:Fungi_of_Saskatchewan.JPG

Taxon Markup

- Taxon profile
- Deployed at National Museum of Natural History of Paris



Examples

Taxon Profile: https://bioschemas.org/profiles/Taxon

Property	Expected Type	Description	CD	Controlled Vocabulary	Example
Marginality: Minimum.					
name	Text	**Schema:** The name of the item. **Bioschemas:** Currently valid (zoological) or accepted (botanical) name for that taxon, with authorship and date information if known.	ONE		
taxonRank	PropertyValue Text URL	**Schema:** The taxonomic rank of this taxon given preferably as a URI from a controlled vocabulary (e.g. the ranks from TDWG TaxonRank ontology or equivalent Wikidata URIs)	MANY		
Marginality: Recommended.					
parentTaxon	Taxon Text URL	**Schema:** Closest parent taxon of the taxon in question. Inverse property: childTaxon **Bioschemas:** Direct, most proximate higher-rank parent taxon	ONE		
url	URL	**Schema:** URL of the item.	ONE		

Bioschemas Taxon Profile

http://bioschemas.org/profiles/Taxon/

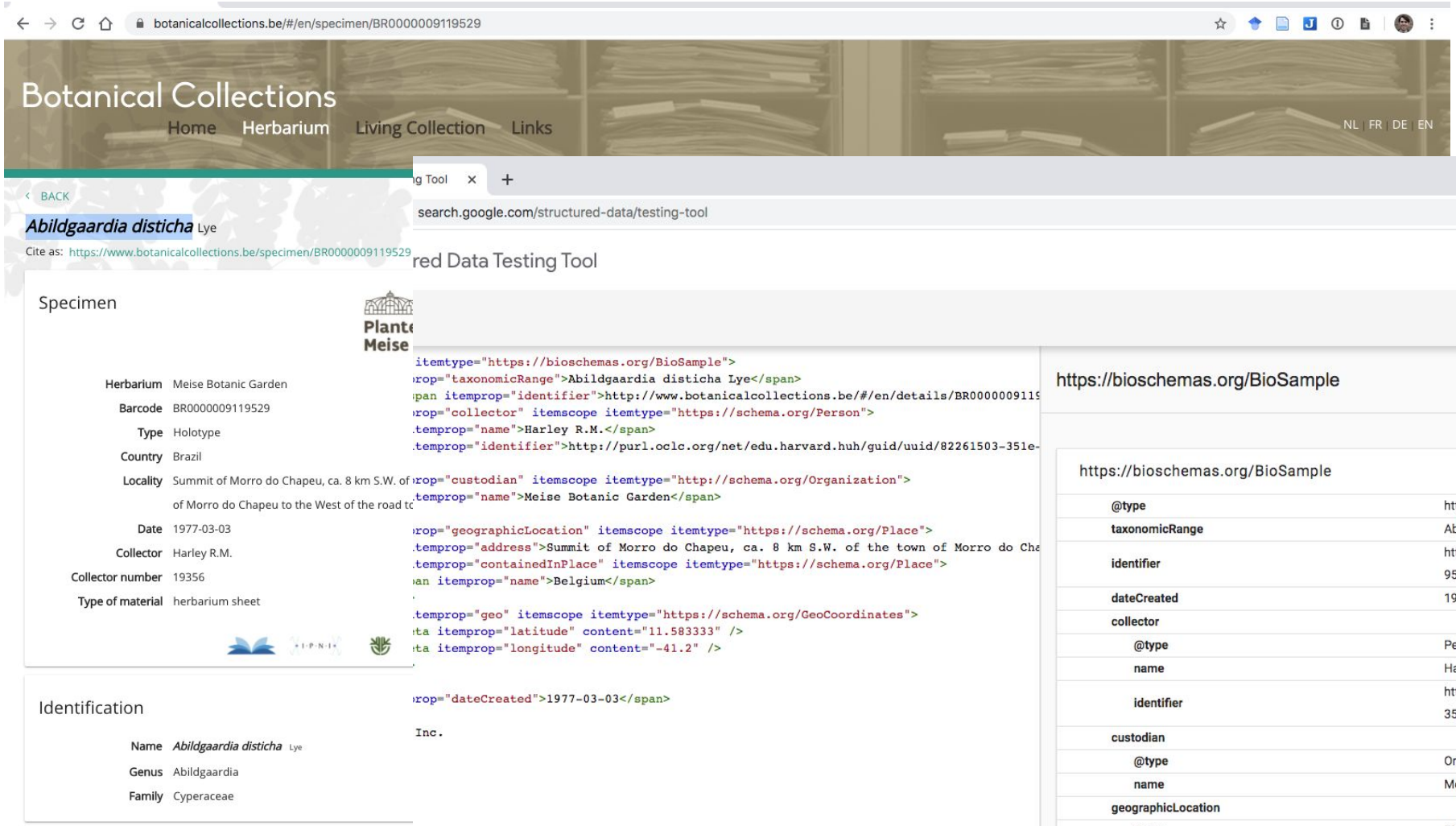
Property	Description	Cardinality
name	Currently valid (zoological) or accepted (botanical) name for that taxon, with authorship and date information if known.	One
taxonRank	The taxonomic rank of this taxon given preferably as a URI from a controlled vocabulary (e.g. the ranks from TDWG TaxonRank ontology or equivalent Wikidata URIs)	Many
parentTaxon	Closest parent taxon of the taxon in question. Inverse property: childTaxon Direct, most proximate higher-rank parent taxon	One
url	Link to the webpage associated to this taxon	One
additionalType	A Taxon type from a well known vocabulary, e.g. DarwinCore http://rs.tdwg.org/dwc/terms/Taxon or http://rs.tdwg.org/ontology/voc/TaxonConcept#TaxonConcept	Many
alternativeName	Scientific name, with authorship and date information if known, of a synonym of the currently valid (zoological) or accepted (botanical) name.	Many
childTaxon	Closest child taxa of the taxon in question. Inverse property: parentTaxon Direct, most proximate lower-rank child taxa	Many
hasCategoryCode	A Category code contained in this code set.	Many
sameAs	URL of third-party webpages describing the same taxon	Many
dwc:vernacularName	A vernacular (common) name of the taxon	Many

BioSample Profile: https://bioschemas.org/profiles/BioSample

Property	Expected Type	Description	CD	Controlled Vocabulary	Example
Marginality: Minimum.					
description	Text	**Schema:** A description of the item.			
identifier	PropertyValue Text URL	**Schema:** The identifier property represents any kind of identifier for any kind of Thing, such as ISBNs, GTIN codes, UUIDs etc. Schema.org provides dedicated properties for representing many of these, either as textual strings or as URL (URI) links. See background notes for more details.			
name	Text	**Schema:** The name of the item.			
url	URL	**Schema:** URL of the item.			
Marginality: Optional.					
additionalProperty	PropertyValue	**Schema:** A property-value pair representing an additional characteristics of the entitity, e.g. a product feature or another characteristic for which there is no matching property in schema.org. Note: Publishers should be aware that applications designed to use specific schema.org properties (e.g. http://schema.org/width, http://schema.org/color, http://schema.org/gtin13, ...) will typically expect such data to be provided using those properties, rather than using the generic property/value mechanism.			
associatedDisease	MedicalCondition	**Schema:**			

BioSample Markup

- BioSample profile
- Deployment in development at Meise Botanical Garden, Belgium



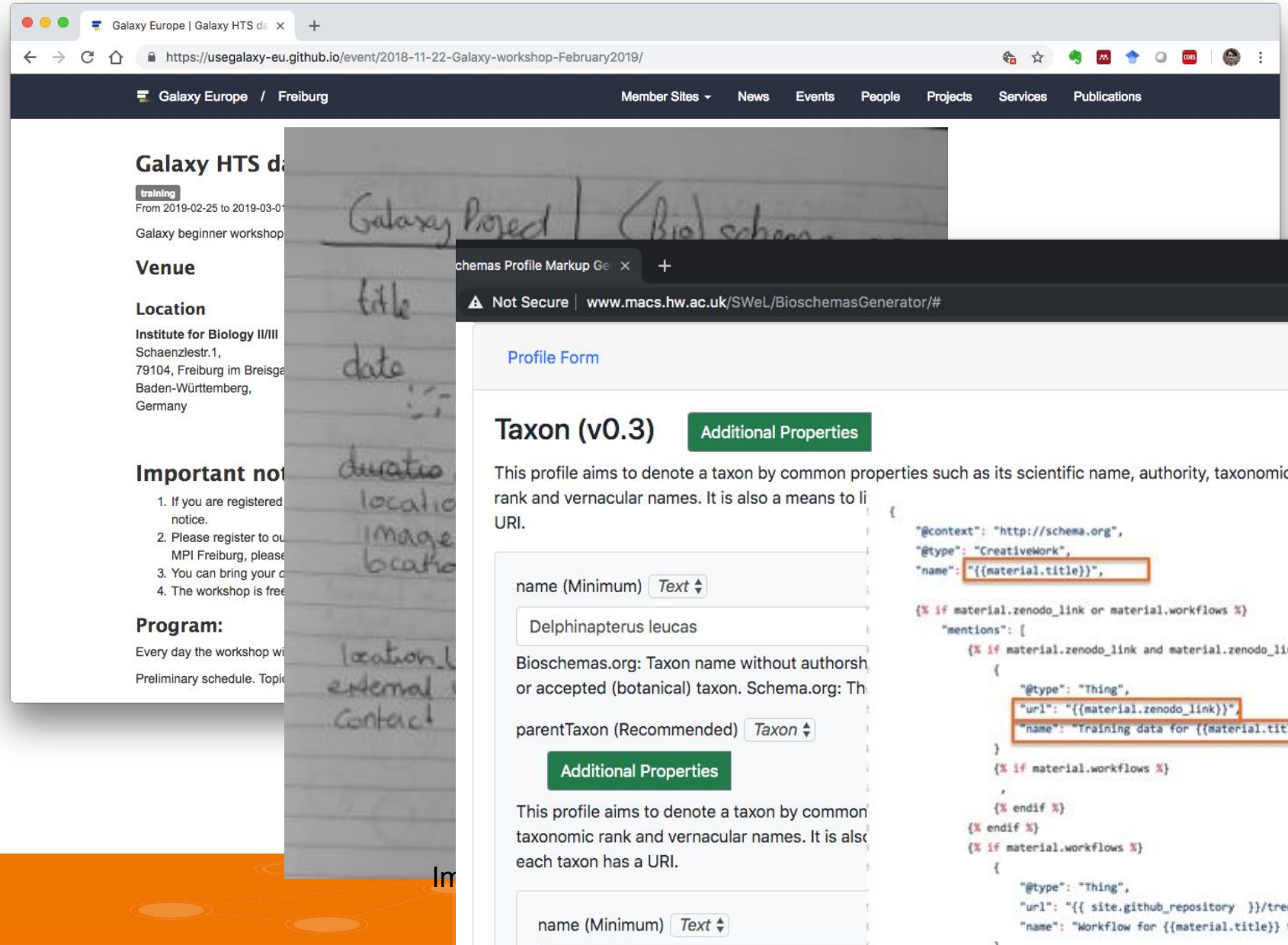
Bioschemas BioSample Profile

http://bioschemas.org/profiles/BioSample/

Property	Description	Cardinality
description	A description of the item.	One
identifier	The identifier property represents any kind of identifier for any kind of Thing, such as ISBNs, GTIN codes, UUIDs etc. Schema.org provides dedicated properties for representing many of these, either as textual strings or as URL (URI) links. See background notes for more details.	One
name	Currently valid (zoological) or accepted (botanical) name for that taxon, with authorship and date information if known.	One
url	Link to the webpage associated to this taxon	One
associatedDisease	Disease associated to this BioChemEntity. Such a disease can be a MedicalCondition or a URL. If you want to add an evidence supporting the association, please use PropertyValue.	Many
bioChemInteraction	A BioChemEntity that is know to interact with this item.	Many
bioChemSimilarity	A similar molecular entity, e.g., obtained by fingerprint similarity algorithm.	Many
biologicalRole	A role played by the molecular entity within a biological context.	Many
collector	The Person or Organization who collected the Sample.	Many
...	...	...

Adding Bioschemas to a Resource

- Identify metadata fields
- Map metadata to Bioschema properties
- Generate JSON-LD
- Embed generation into web server



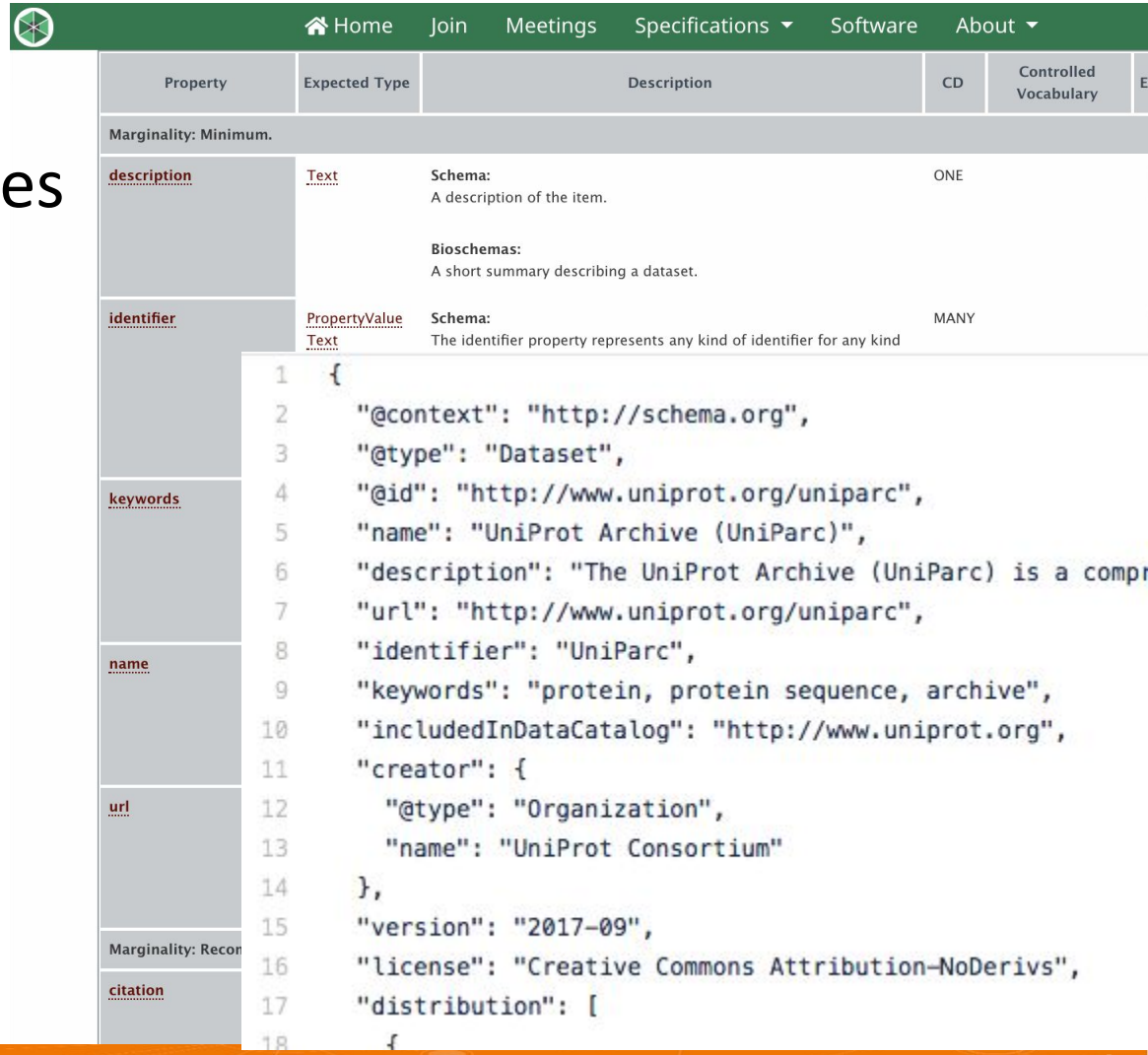
Bioschemas

What?

- Exploiting schema.org to make Life Sciences resources more discoverable
  - Search engines will index and understand markup

How?

- Extending schema.org vocabulary for life sciences
- Provide guidelines on how to markup resources
- Biodiversity types and profiles available




Bioschemas Community
30
Profiles
89
Sites
bioschemas.org/
liveDeploys
14+
Countries
20M+
Pages
bioschemas.org/liveDeploys
300+
People
17
Types

Acknowledgements http://bioschemas.org/people

- Ahmed Ali
- Stephen Anyango
- Ricardo Arcila
- Haydee Artaza
- Mohammad Asif Khan
- Christian Atallah
- Teresa Attwood
- Phil Barker
- Dominique Batista
- Michael Baudis
- Roman Baum
- Niall Beard
- Jerven Bolleman
- Michelle Brazas
- Cath Brooksbank
- Tony Burdett
- Guillermo Calderon Mantilla
- Ethy Cannon
- Denise Carvalho-Silva
- Luca Cherubin
- Justin Clark-Casey
- Dave Clements
- Martin Cook
- Manuel Corpas
- Melanie Courtot
- Michael R Crusoe
- Pavel Dallakian
- Luc Deltombe
- Victoria Dominguez Del Angel
- Michel Dumontier
- Nathan Dunn
- Stephen Ficklin
- Robert Finn
- Alexander Garcia
- Leyla Garcia
- Anna Gaulton
- Carole Goble
- Alejandra Gonzalez-Beltran
- Alasdair Gray
- Matthew Green
- Jeffrey Grethe
- Quentin Groom
- Henning Hermjakob
- Petr Holb
- Richard Holland
- Carlos Horro
- Jon Ison
- Christa Janko
- Andy Jenkinson
- Rafael C Jimenez
- Claire Johnson
- Simon Jupp
- Nick Juty
- Jasper Koehorst
- Eija Korpelainen
- Mateusz Kuzak
- Lee Larcombe
- Yvan Le Bras
- Nicolas Le Novère
- Mikael Linden
- Audald Lloret
- Ankit Kumar Lohani
- Aurélien Luciani
- Federico López Gómez
- Ronald Margolis
- Maria Martin
- Michaela Th. Mayrhofer
- Annette McGrath
- James Alastair McLaughlin
- Kenneth McLeod
- Peter McQuilton
- David Mendez
- Ann Meyer
- Franck Michel
- Gos Micklem
- Saqib Mir
- Sarah Morgan
- Chris Mungall
- Heimo Müller
- Aleksandra Nenadic
- Michał Nowotka
- Francis Ouellette
- Stuart Owen
- Petros Papadopoulos
- Helen Parkinson
- Roberto Preste
- Giuseppe Profiti
- Verena Ras
- Dietrich Rebholz-Schuhmann
- Anders Riutta
- Philippe Rocca-Serra
- Gabriella Rustici
- Gustavo Salazar-Orejuela
- Sheeba Samuel
- Susanna A Sansone
- Vicky Schneider
- Serena Scollen
- Kaisa Silander
- Stian Soiland-Reyes
- Morris Swertz
- Mohameth François Sy
- Chris Taylor
- Milo Thurston
- Dan Timmons
- Sonika Tyagi
- John Van Horn
- Susheel Varma
- Sameer Velankar
- Premysl Velek
- Andra Waagmeester
- Alan Williams
- Liz Williams
- Egon Willighagen
- Sarala Wimalaratne
- Anil Wipat
- Chunlei Wu
- Olga Ximena Giraldo
- Matt Yoder
- Gianluigi Zanetti
- Charlotte Zwetsloot
- Anita de Waard
- David van Enckevort
- Celia van Gelder
- Peter van Heusden

bioschemas.org

@bioschemas

github.com/bioschemas

Special thanks to:

- Franck Michel
- Quentin Groom
- Matt Styles

Join Bioschemas: http://bioschemas.org/howtojoin/