



Harvesting Bioschemas

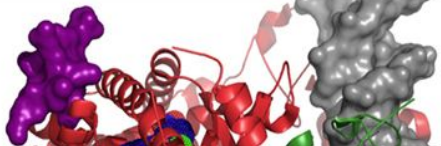
Markup

Alasdair Gray

SWAT4HCLS Tutorial
10 January 2022

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

Bioschemas: Markup for the Life Sciences

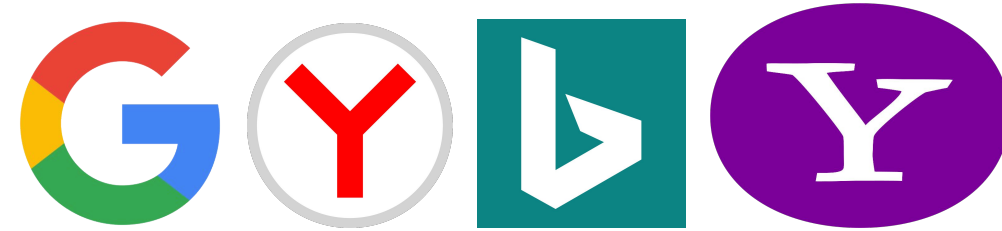
 IUPHAR/BPS
Guide to PHARMACOLOGY

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-CoV infection. PMID: 28553736
ACE2 ligands		
anakinra		
anti-TNF therapy (e.g. infliximab and others)		
apilimod		
ASC09F		

```
<!-- Bioschemas -->
<script type="application/ld+json">[
{
  "@context": "https://schema.org",
  "@type": "Protein",
  "@id": "https://www.guidetopharmacology.org/GRAC/ObjectDisplayForward?objectId=1614",
  "http://purl.org/dc/terms/conformsTo": "https://bioschemas.org/profiles/Protein/0.4-DRAFT-2019_05_20/",
  "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": "acalabrutinib",
  "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 of..."
}
```



orphanet



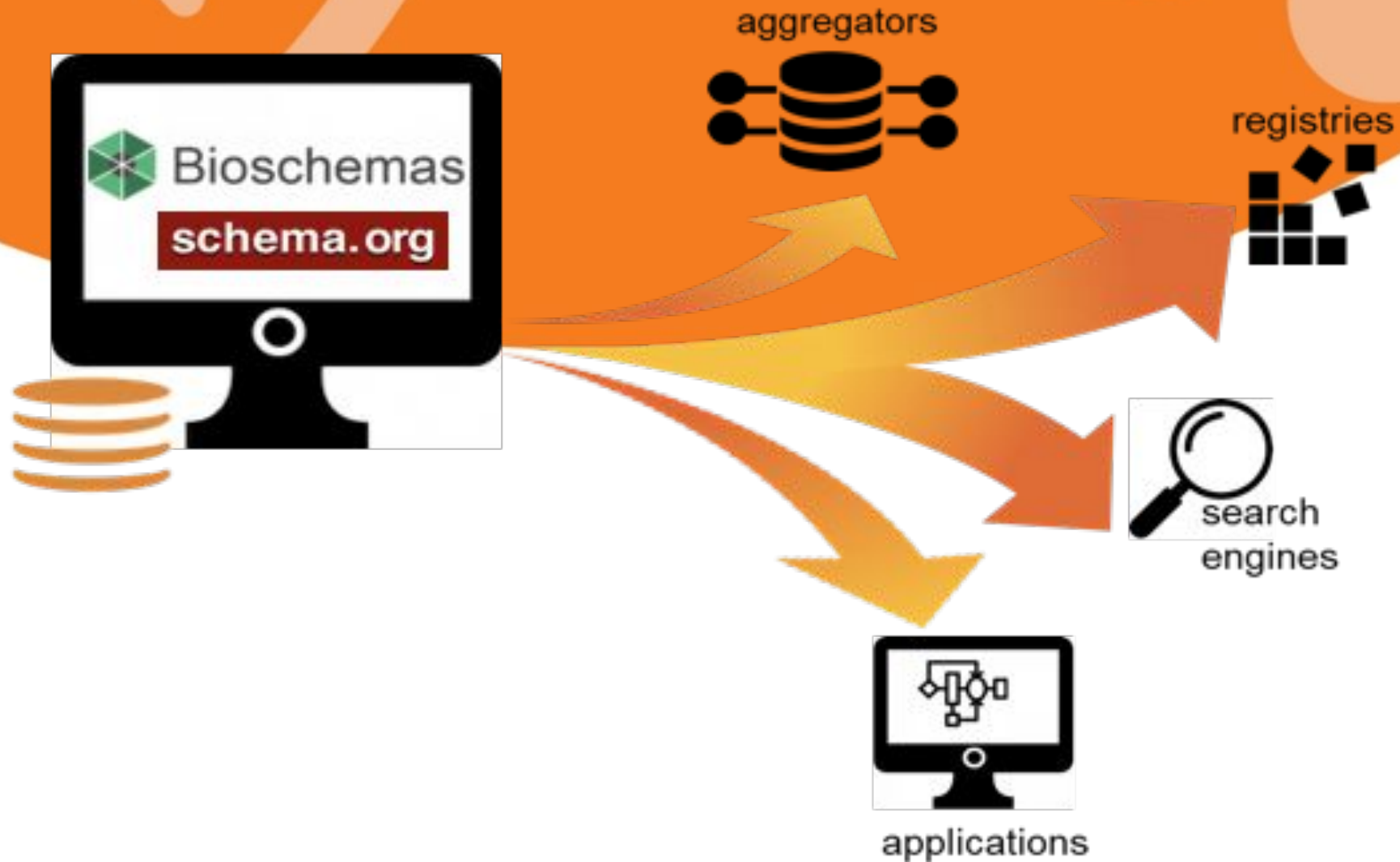
The DataCite
PID Graph



http://researchgraph.org/



Exploiting Bioschemas Markup



Exploiting Bioschemas Markup

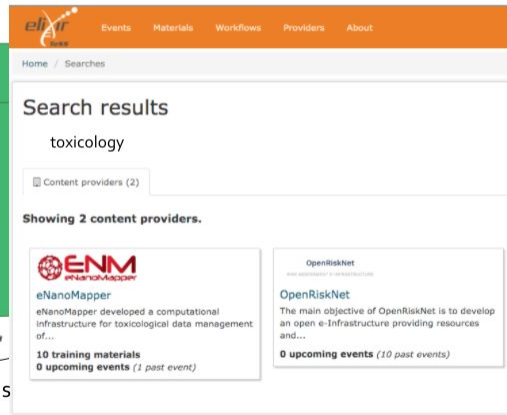
Specialised Search: TeSS

Bioschemas Course:

Bioschemas CourseInstance:

Bioschemas TrainingMaterial:

- description
- keywords
- name
- about
- abstract

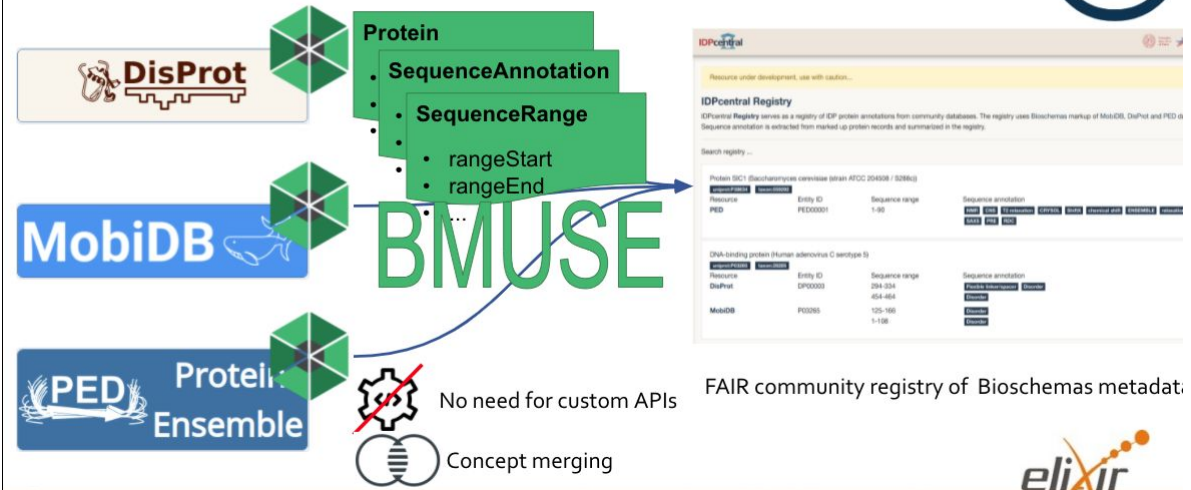


- No need for custom APIs
- No concept merging



- Concept merging
- Terminology transformation
- No need for custom APIs

Community Registry: IDPcentral



FAIR community registry of Bioschemas metadata



{B}EH Project 29: Bioschemas

Local

bioschemas

total statements
11,839,960

Sites harvested
25 (partially)

Pages harvested
413,748

<https://swel.macs.hw.ac.uk/data/>

Query: Cou



Specialised Search: TeSS

BiData.pt
Portuguese Biological Data Network

DMOS18 Data Management and C

Course Description
In an age of increasingly complex and data-intensive, collaborative scientific practices, training and accountability, a new paradigm has arisen: Open Science. In this four day long course, we will explore the paradigm (Open Access scholarly publishing, Open Source software development, Open Data integration and sharing). For this, we will be introducing a set of technologies and ways of using them. The reasonable use of these tools for the above purposes is a highly productive way. The use cases that we will be exploring are very broadly applicable to other fields. You do not need to have any particular background or training from a scientist to graduate school and beyond. I.e., you should not be afraid. We will then show the participants the good practices in data management and such as project, these practices influence reproducibility and reusability. At the end we will discuss publication strategies. You can see the detailed program through the link below.
Detailed Program
Registration
Registration will be open until June 15th, following the instructions in the link below:
Registration instructions
Contact: For any questions regarding this course, please contact OTPE (contact below).
Start date:
2018-06-12
End date:

Metabolomics with the Q Exactive mass spectrometer
Register Now for 19-21 November 2018

Overview
This 3-day course introduces you to using the Q Exactive mass spectrometer. The course is led by experts in the field of metabolomics and includes workshops to provide a detailed overview of the metabolomics pipeline applying to metabolomics.

UPCOMING EVENTS (CO-ORGANIZED BY IFB)
June 2018
19th May de novo assembly - Sarsibar (second session) - The SB-OS
Training course for bioinformaticians and biologists aiming at introducing a new approach.
20-24 May 18 (7 days) - Inra, INRA de Bordeaux, Aquitaine
Registration deadline: 2017-12-18
Link to event: <http://events.genouest.fr/index.php/event/19th-de-novo-assembly>
11-12 June 2018 - The SB-OS
CURIO organise une formation sur la direction de l'annotation des éléments transcrits par un pipeline RPKM
Transcription par un pipeline RPKM
From 11-11 to 12-12 2018 (2 days) - INRA de Versailles
Registration deadline: Not available
Link to event: <http://events.genouest.fr/index.php/event/11-12-june-2018>

Upcoming training events
RNA-seq: From quality control to pathway analysis
20-21 May 2018
First Steps in Parallelization with OpenMP
22-23 May 2018

Bioschemas Course:

- Bioschemas CourseInstance:

- Bioschemas TrainingMaterial:

- description
- keywords
- name
- about
- abstract

TeSS



No need for custom APIs



No concept merging

elixir TeSS

Events Materials Workflows Providers About

Home / Searches

Search results

toxicology

Content providers (2)

Showing 2 content providers.

eNanoMapper
eNanoMapper developed a computational infrastructure for toxicological data management of...

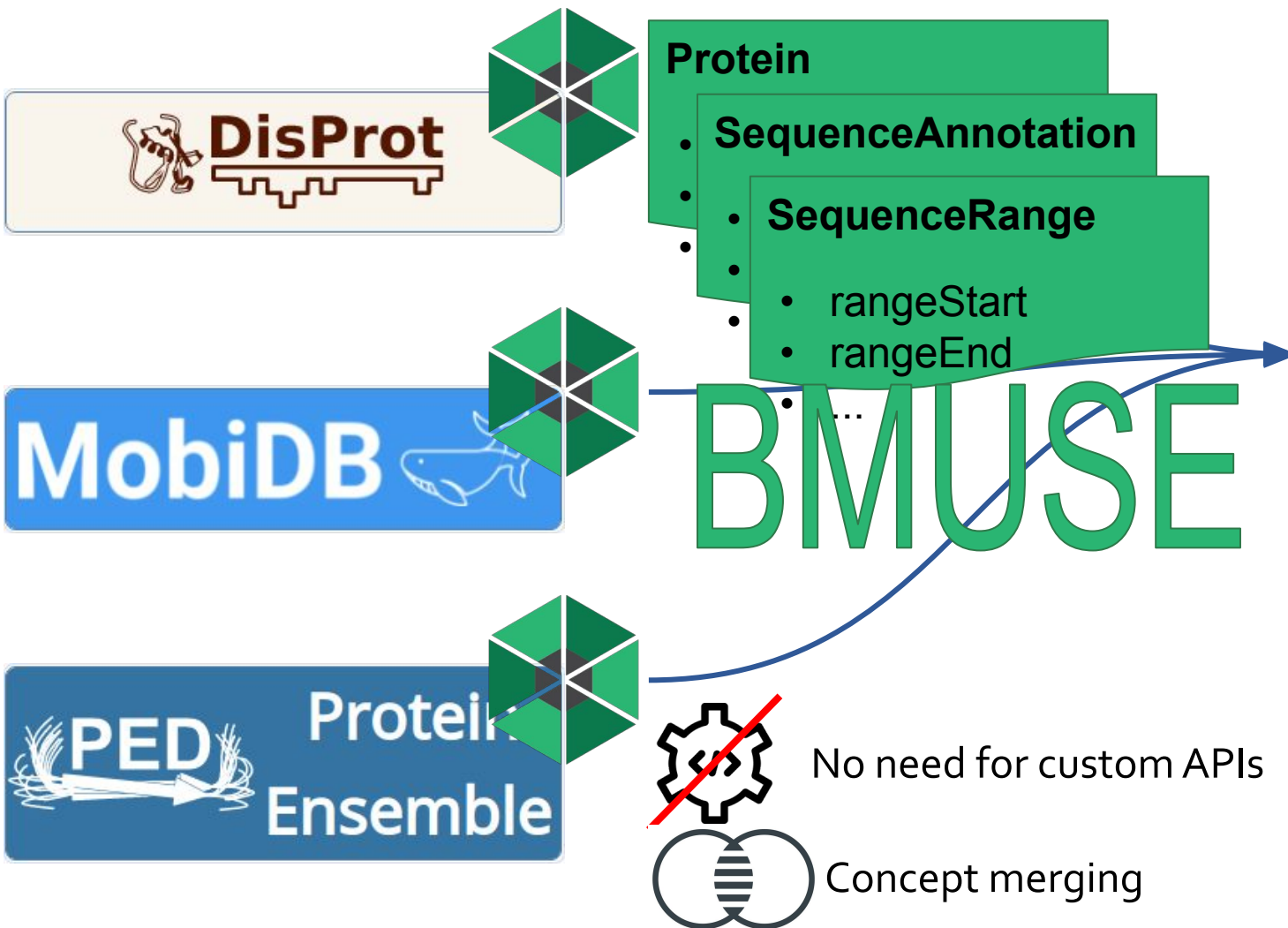
10 training materials
0 upcoming events (1 past event)

OpenRiskNet
RISK ASSESSMENT E-INFRASTRUCTURE

OpenRiskNet
The main objective of OpenRiskNet is to develop an open e-Infrastructure providing resources and...

0 upcoming events (10 past events)

Community Registry: IDPcentral



Resource under development, use with caution...

IDPcentral Registry

IDPcentral **Registry** serves as a registry of IDP protein annotations from community databases. The registry uses Bioschemas markup of MobiDB, DisProt and PED data. Sequence annotation is extracted from marked up protein records and summarized in the registry.

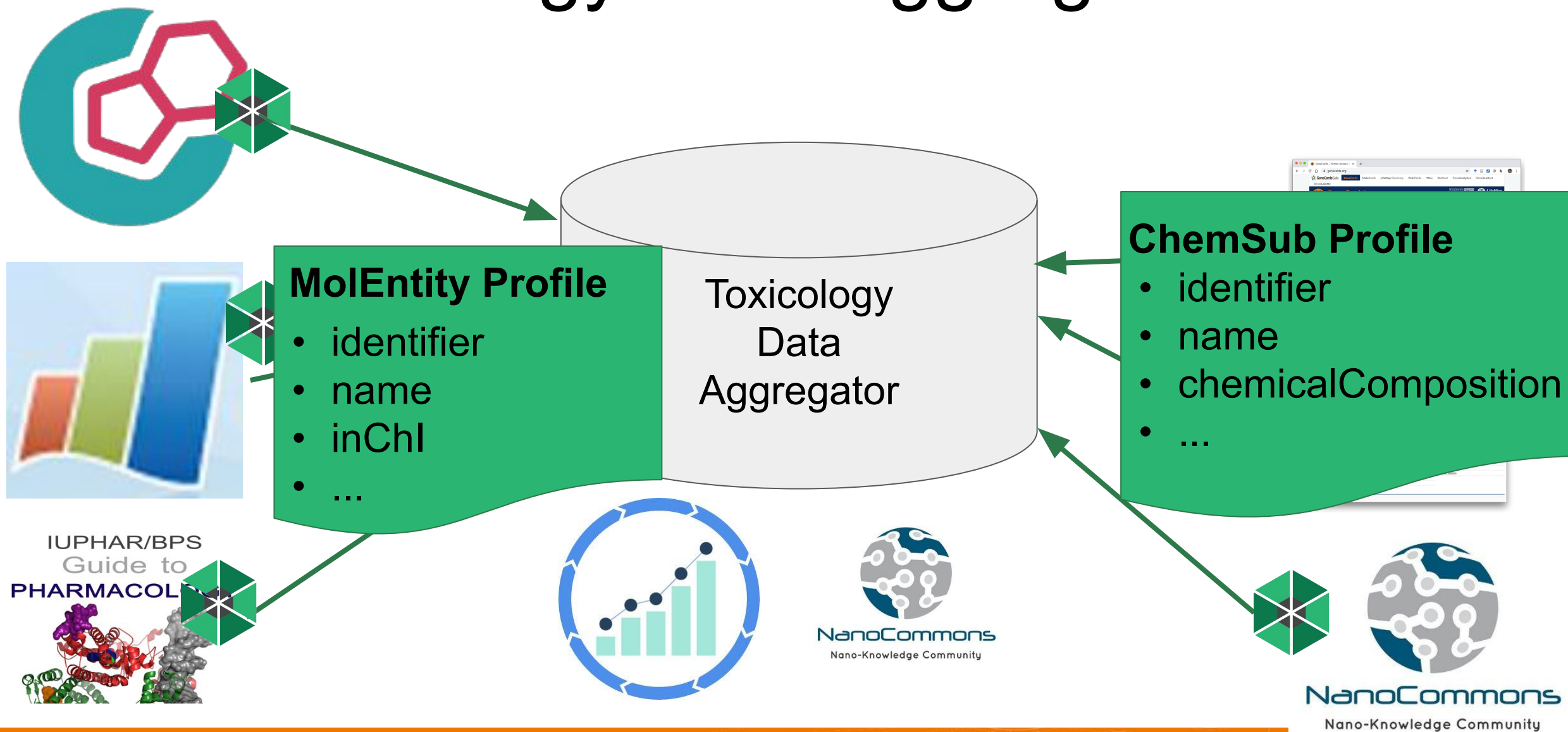
Search registry ...

Protein	uniprot	taxon	Resource	Entity ID	Sequence range	Sequence annotation
SIC1 (Saccharomyces cerevisiae (strain ATCC 204508 / S288c))	uniprot:P38634	taxon:559292	PED	PED00001	1-90	NMR CNS T2 relaxation CRYSOLO ShiftX chemical shift ENSEMBLE relaxation SAXS PRE RDC
DNA-binding protein (Human adenovirus C serotype 5)	uniprot:P03265	taxon:28285	DisProt	DP00003	294-334 454-464	Flexible linker/spacer Disorder
	MobiDB	P03265			125-166 1-108	Disorder Disorder

FAIR community registry of Bioschemas metadata



Toxicology Data Aggregation



{B|H} Project 29: Bioschemas

Local

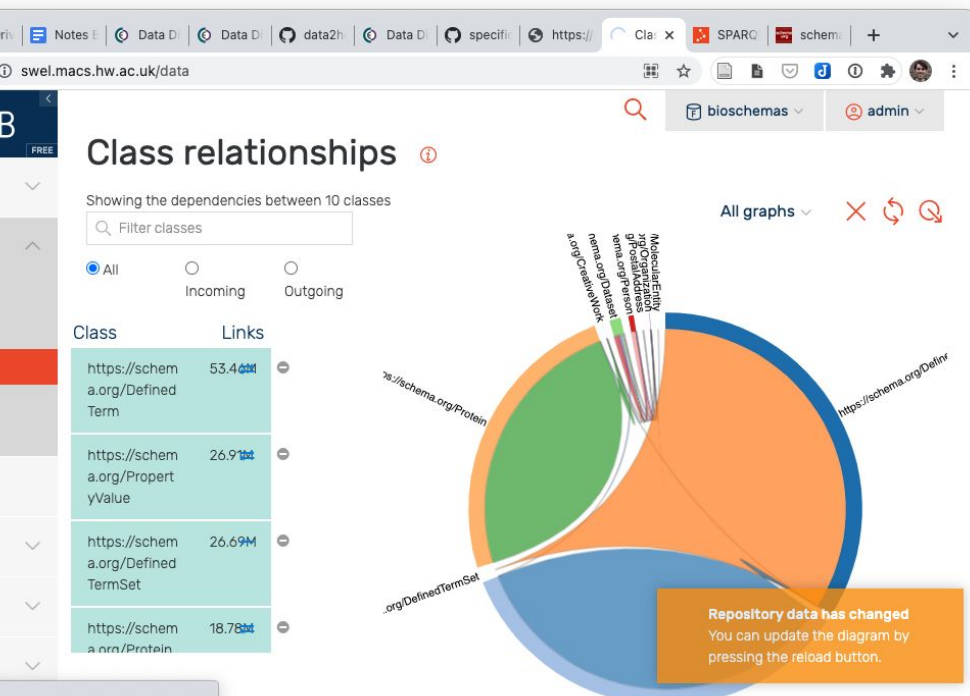
 bioschemas

total statements
11,839,960

Sites harvested
25 (partially)

Pages harvested
413,748

<https://swel.macs.hw.ac.uk/data/>



AnalysisQueries.ipynb

Python 3 (ipykernel)

[6]: createSelectionGUI()

Instances per Bioschemas Type

Execute

Query: Instances per Bioschemas Type (File: /queries/queries/instancesPerBioschemasType.rq)

Results

Copy .md

Number of results: 7

Class	instances
https://bioschemas.org/Gene	238079
https://schema.org/Dataset	201302
https://schema.org/MolecularEntity	199350
https://bioschemas.org/Taxon	55884
https://schema.org/ScholarlyArticle	9350
https://schema.org/Protein	4462
https://schema.org/Gene	39

Query: Cou

Results

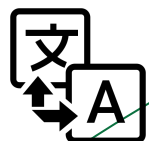
Number of results: 146

domain	type	count
www.gbif.org	https://schema.org/PostalAddress	297090
www.gbif.org	https://schema.org/Person	291260
bgee.org	https://bioschemas.org/Gene	263793
www.gbif.org	https://schema.org/Organization	186688
www.gbif.org	https://schema.org/PropertyValue	126268
massbank.eu	https://schema.org/Dataset	76249
massbank.eu	https://schema.org/MolecularEntity	76249

Global Research Graph: OpenAIRE



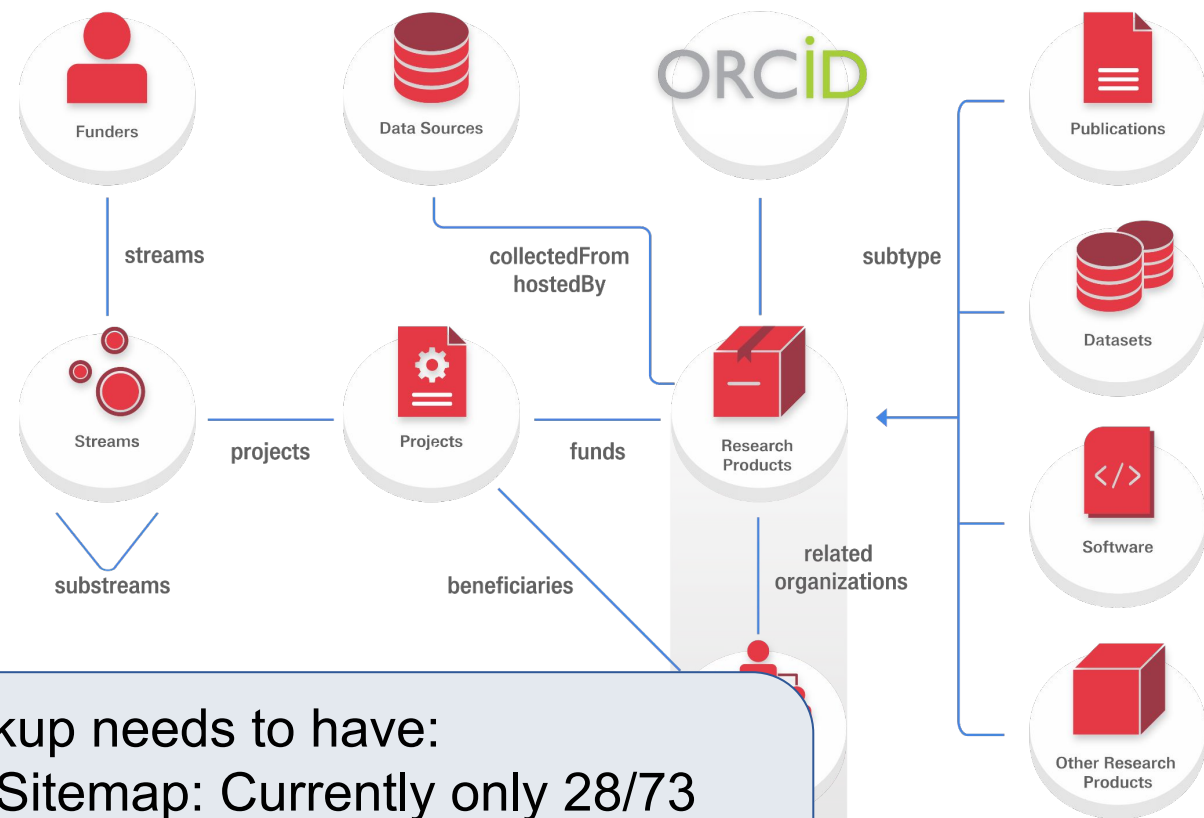
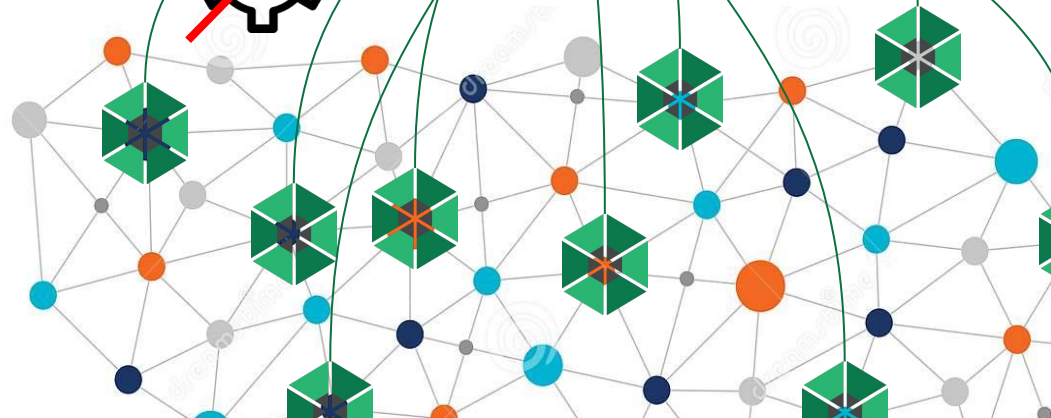
Concept merging



Terminology transformation



No need for custom APIs



Markup needs to have:

- Sitemap: Currently only 28/73
- Link to
 - Publication
 - Dataset



Project 29: Bioschemas



- Deploy markup
 - Include links to publications
- Implement sitemap
- Register on FAIRsharing
- Harvest data
- Map Bioschemas to Datacite
- Harvest data
- Align identifiers
- Store in KG

BioSchema specifications for harvesting

File Edit View Insert Format Data Tools Extensions Help Last edit was made yesterday at 15:55 by A.C.z

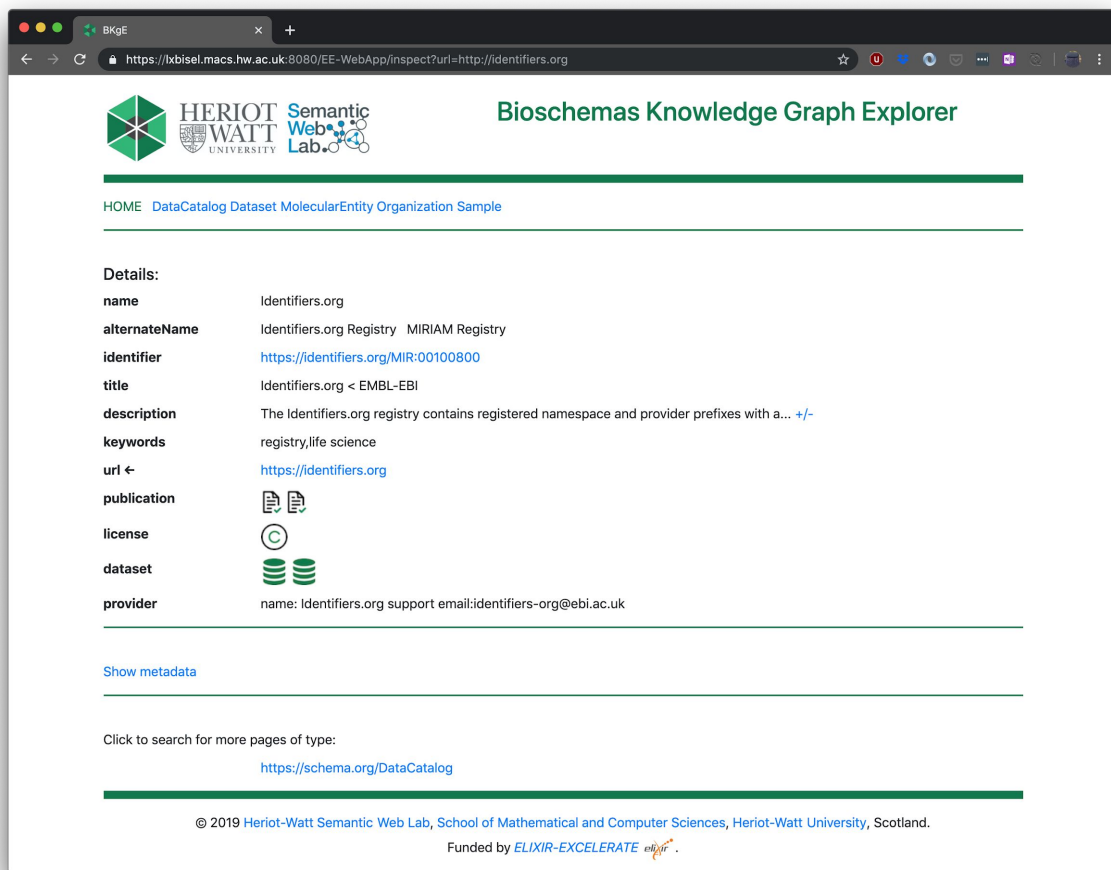
100% 123 arial 14

Gene BioSchema v1.1-Draft

A	J	K	L	M	N
1	Gene BioSchema v1.1-Draft				
2					
3	schema.org / pending / External ontologies				
4	Property	Example	Appl. Profile Elements (from inst.&them. Guidelines)	req. level	Card.
5	@type	"@type": "Gene"	caire:resourceType	M	1 set fix
6	alternateName	{ "@type": "Gene", "alternateName": "AD1" }	titleType="AlternateTitle" , xml:lang	M if alternateId	1 will skip in 1st row
7	alternativeOf				
8	associatedDisease		datacite:relatedIdentifier	R	0-n
9	description	{ "@type": "Gene", "description": "amyloid beta precursor" }	datacite:description	MA	0-n

BKG Explorer

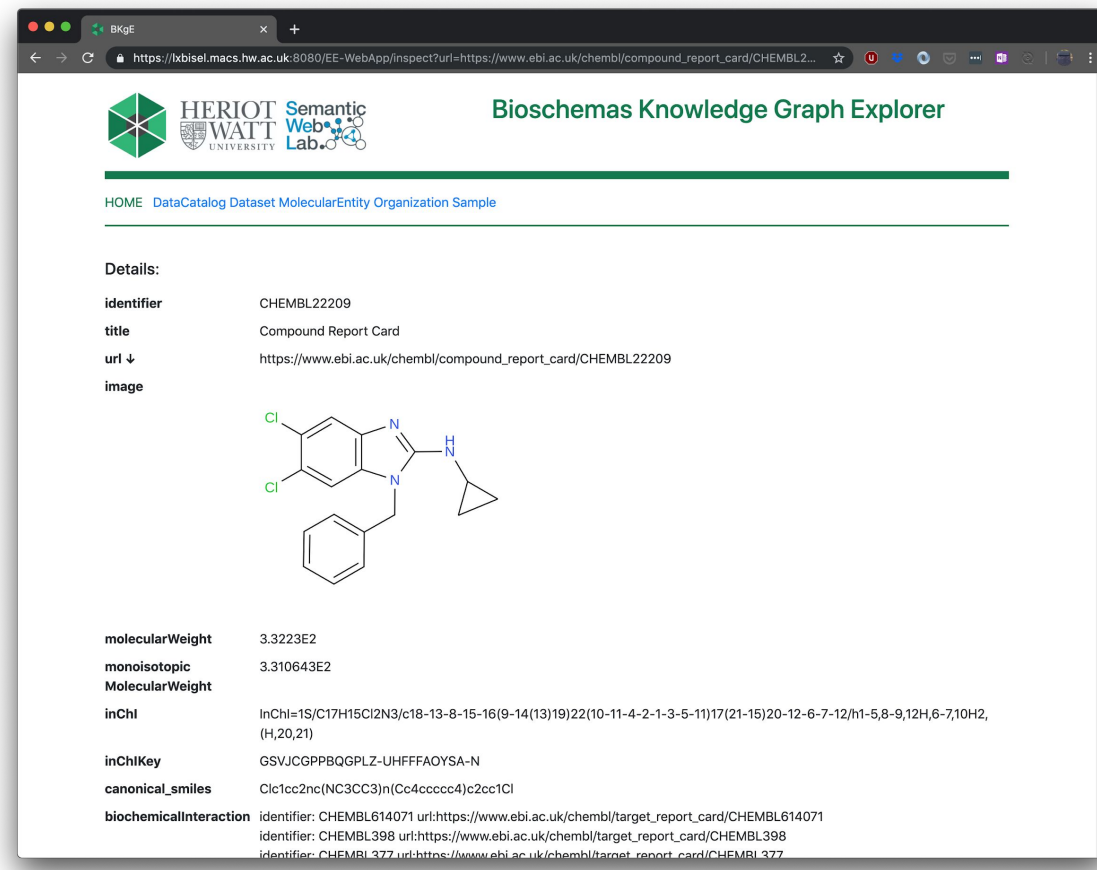
Built over Bioschemas markup crawled from 30 live deployments
20,000 pages



The screenshot shows the BKG Explorer interface for the Identifiers.org dataset. The header includes the Heriot-Watt University Semantic Web Lab logo and the title "Bioschemas Knowledge Graph Explorer". A navigation bar contains links: HOME, DataCatalog, Dataset, MolecularEntity, and Organization Sample. The main content area is titled "Details:" and lists the following information:

- name:** Identifiers.org
- alternateName:** Identifiers.org Registry MIRIAM Registry
- identifier:** <https://identifiers.org/MIR-00100800>
- title:** Identifiers.org < EMBL-EBI
- description:** The Identifiers.org registry contains registered namespace and provider prefixes with a... +/-
- keywords:** registry, life science
- url:** <https://identifiers.org>
- publication:** (represented by document icons)
- license:** (represented by a Creative Commons icon)
- dataset:** (represented by a database icon)
- provider:** name: Identifiers.org support email: identifiers-org@ebi.ac.uk

Below the details, there is a "Show metadata" link and a search prompt: "Click to search for more pages of type: <https://schema.org/DataCatalog>". The footer contains the copyright notice: "© 2019 Heriot-Watt Semantic Web Lab, School of Mathematical and Computer Sciences, Heriot-Watt University, Scotland." and the funding source: "Funded by ELIXIR-EXCELERATE".



The screenshot shows the BKG Explorer interface for the ChEMBL22209 compound report card. The header and navigation bar are identical to the previous screenshot. The main content area is titled "Details:" and lists the following information:

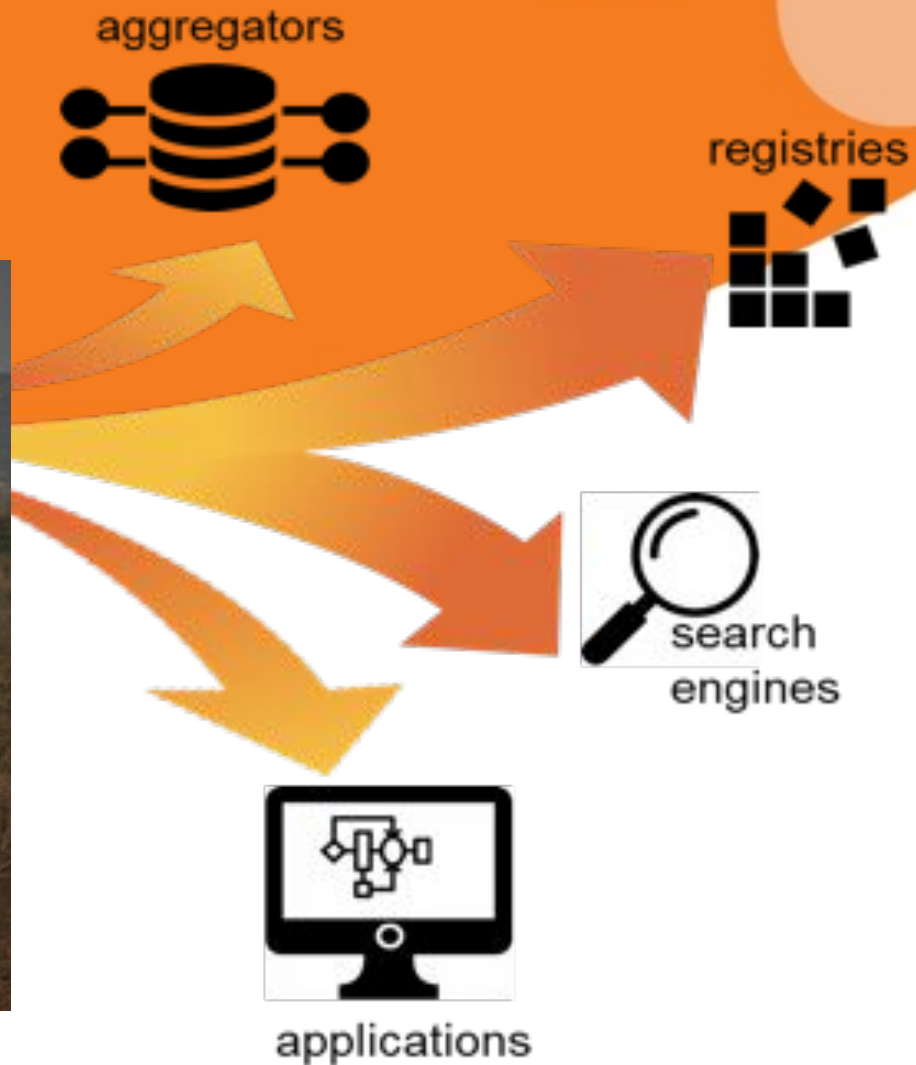
- identifier:** ChEMBL22209
- title:** Compound Report Card
- url:** https://www.ebi.ac.uk/chembl/compound_report_card/ChEMBL22209
- image:** (represented by a chemical structure image)

The chemical structure image shows a complex molecule with a benzene ring, a nitrogen atom, and a cyclopropyl group. Below the image, the following information is listed:

- molecularWeight:** 3.3223E2
- monoisotopic MolecularWeight:** 3.310643E2
- inChI:** InChI=1S/C17H15Cl2N3/c18-13-8-15-16(9-14(13)19)22(10-11-4-2-1-3-5-11)17(21-15)20-12-6-7-12/h1-5,8-9,12H,6-7,10H2,(H,20,21)
- inChIKey:** GSVJCGPPBQGPLZ-UHFFFAOYSA-N
- canonical_smiles:** Clc1cc2nc(NC3CC3)n(Cc4ccccc4)c2cc1Cl
- biochemicalInteraction:** identifier: ChEMBL614071 url: https://www.ebi.ac.uk/chembl/target_report_card/ChEMBL614071
identifier: ChEMBL398 url: https://www.ebi.ac.uk/chembl/target_report_card/ChEMBL398
identifier: ChEMBL377 url: https://www.ebi.ac.uk/chembl/target_report_card/ChEMBL377



Harvesting Bioschemas Markup



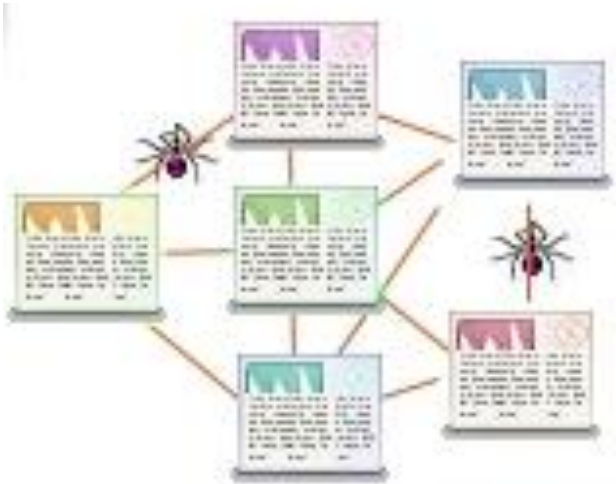
Crawling or Harvesting

Crawling

- Seed list of URLs
- Pages retrieved
- URLs extracted and added to seed list
- Data extracted from page
- Never ending

Harvesting

- Seed list of URLs or sitemaps
- Pages retrieved
- Data extracted from page
- Finite set of pages extracted



Static or Dynamic Sites

Static

- Markup in the HTML provided by the server
- Markup visible in viewsource
- No client side processing

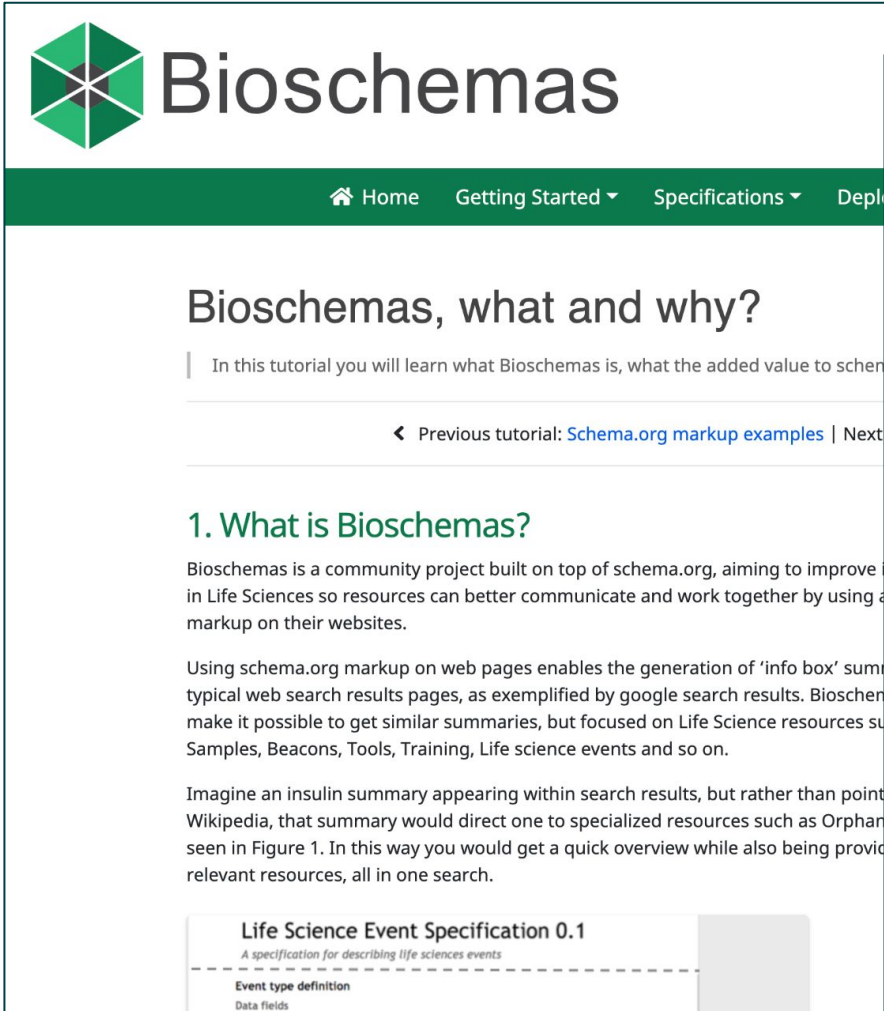
Dynamic

- Javascript Single page application
- Client side/browser calls to retrieve pages
- Rendered page not in 'viewsource'
- Larger processing overhead



Static or Dynamic Sites

https://bioschemas.org/tutorials/what_why_bioschemas



The screenshot shows the Bioschemas website header with the logo and navigation links: Home, Getting Started, Specifications, and Deploy. The main heading is 'Bioschemas, what and why?'. Below it, a subheading reads 'In this tutorial you will learn what Bioschemas is, what the added value to schema.org is and what the main elements in Bioschemas are'. A link points to the 'Previous tutorial: Schema.org markup examples'. The first section is '1. What is Bioschemas?'. The text explains that Bioschemas is a community project built on top of schema.org, aiming to improve life sciences resources by using schema.org markup. It mentions that using schema.org markup enables the generation of 'info box' summaries in typical web search results, as exemplified by Google search results. It states that Bioschemas makes it possible to get similar summaries, but focused on life science resources such as Samples, Beacons, Tools, Training, Life science events and so on. An example is given: 'Imagine an insulin summary appearing within search results, but rather than pointing to Wikipedia, that summary would direct one to specialized resources such as Orphanet'. The bottom of the screenshot shows a table titled 'Life Science Event Specification 0.1' with columns for 'Event type definition' and 'Data fields'.

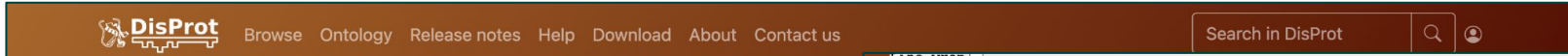
```
Line wrap
1 <!DOCTYPE html>
2 <html lang="en">
3 <html>
4 <head>
5 <meta charset="utf-8">
6 <meta http-equiv="X-UA-Compatible" content="chrome=1">
7 <meta name="viewport" content="width=device-width, initial-scale=1">
8 <meta name="msapplication-TileColor" content="#da532c">
9 <meta name="theme-color" content="#ffffff">
10 <meta property="og:image" content="//bioschemas.org/assets/img/apple-touch-icon.png" />
11 <title>
12   Bioschemas - Bioschemas, what and why?
13 </title>
14 <meta name="description" content="1. What is Bioschemas?">
15 <!-- Icons -->
16 <link rel="apple-touch-icon" sizes="180x180" href="/assets/img/apple-touch-icon.png">
17 <link rel="icon" type="image/png" sizes="32x32" href="/assets/img/favicon-32x32.png">
18 <link rel="icon" type="image/png" sizes="16x16" href="/assets/img/favicon-16x16.png">
19 <link rel="manifest" href="/assets/img/site.webmanifest">
20 <link rel="mask-icon" href="/assets/img/safari-pinned-tab.svg" color="#0b794b">
21 <link rel="shortcut icon" href="">
22 <!-- Styling -->
23 <link href="https://fonts.googleapis.com/css?family=Noto+Sans" rel="stylesheet" type="text/css">
24 <link rel="stylesheet" href="/assets/css/main.css">
25 <link rel="canonical" href="https://bioschemas.org/tutorials/what_why_bioschemas">
26 <link rel="alternate" type="application/rss+xml" title="Bioschemas" href="/feed.xml">
27
28 <script type="application/ld+json">
29   {
30     "@context": "https://schema.org/",
31     "@type": "LearningResource",
32     "http://purl.org/dc/terms/conformsTo":
33       [
34         {
35           "@type": "CreativeWork",
36           "id": "https://bioschemas.org/profiles/TrainingMaterial/0.9-DRAFT-2020_12_08/",
37           "about": {
38             "@id": "https://schema.org/",
39             "@type": "CreativeWork",
40             "id": "https://bioschemas.org/profiles/TrainingMaterial/0.9-DRAFT-2020_12_08/",
41             "audience": {
42               "@type": "Audience",
43               "name": "(General interest) People interested in introductory information to Bioschemas"
44             },
45             "name": "Bioschemas, what and why?",
46             "author": {
47               "@type": "Person",
48               "name": "Leyla Garcia",
49               "id": "https://bioschemas.org/people/LeylaGarcia",
50               "url": "https://bioschemas.org/people/LeylaGarcia"
51             },
52             "contributor": {
53               "@type": "Person",
54               "name": "Ricardo Arcila",
55               "id": "https://bioschemas.org/people/RicardoArcila",
56               "url": "https://bioschemas.org/people/RicardoArcila"
57             },
58             "id": "https://bioschemas.org/people/VictoriaDominguezDelAngel/",
59             "url": "https://bioschemas.org/people/VictoriaDominguezDelAngel/"
60           },
61           "dateModified": "2021-02-17",
62           "description": "In this tutorial you will learn what Bioschemas is, what the added value to schema.org is and what the main elements in Bioschemas are",
63           "keywords": "schemaorg, markup, structured data, bioschemas",
64           "license": "CC-BY 4.0",
65           "version": "2.0"
66         }
67       ]
68   }
69 </script>
70
71 <script src="https://ajax.googleapis.com/ajax/libs/jquery/3.6.0/jquery.min.js"></script>
72 <script src="/assets/js/bootstrap.bundle.min.js"></script>
73 <script src="/assets/js/toc.js"></script>
74 <script defer src="/assets/js/fontawesome-all.js"></script>
75 </head>
76
77 <body>
78   <a class="visually-hidden-focusable" href="#main">Skip to content</a>
79 <header id="header_wrap" class="p-3">
80   <a class="btn btn-github rounded-0 rounded-bottom shadow text-white fs-5 d-flex align-items-center" href="https://github.com/Bioschemas">
```

view-source:https://bioschemas.org/tutorials/what_why_bioschemas



Static or Dynamic Sites

<https://disprot.org/DP00086>



DP00086 - Cellular tumor antigen p53

Organism [Homo sapiens](#) Sequence length 393 Disorder content 29.3%

Cross references [MobiDB:P04637](#), [UniProtKB:P04637](#), [PED: P04637](#), [FuzDB: FC00084](#)

Last update 2021-12-06

[Collapse Feature-Viewer](#) [Expand Feature-Viewer](#)

DisProt consensus

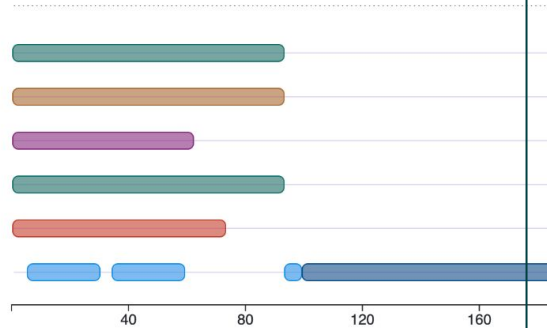
> Structural state

> Structural transition

> Interaction partner

> Disorder function

> Domains



Filter regions

Example filters [structural state](#) [transition](#) [interaction partner](#) [disorder function](#) [magnetic resonance](#)

DP00086r077 Structural state evidence

```
Line wrap
1 <!doctype html>
2 <html lang="en">
3 <head>
4   <meta charset="utf-8">
5   <title>DisProt</title>
6   <base href="/">
7   <meta name="viewport" content="width=device-width, initial-scale=1">
8   <link rel="icon" type="image/x-icon" href="favicon.ico">
9   <!-- Google Analytics -->
10  <script>
11    window.ga = window.ga || function () {
12      (ga.q = ga.q || []).push(arguments)
13    };
14    ga.l = +new Date;
15    ga('create', 'UA-84751070-1', 'auto');
16    ga('send', 'pageview');
17  </script>
18  <script async="" src="https://www.google-analytics.com/analytics.js"></script>
19  <!-- End Google Analytics -->
20  <script type="text/javascript" src="https://apicuron.org/assets/apicuron-widget.js"></script>
21
22  <script src="https://platform.twitter.com/widgets.js" charset="utf-8"></script>
23  <link rel="stylesheet" href="styles.e5a65970d00f70a296e7.css"></head>
24  <body>
25    <app-root></app-root>
26    <script src="runtime.7b63b9fd40098a2e8207.js" defer></script><script src="polyfills-es5.3d81abf619c8b35d02f2.js" nomodule defer></script><script src="polyfills.e4b6aa3527370197aafb">
27  </html>
28
```

view-source:<https://disprot.org/DP00086>



Static or Dynamic Sites

<https://disprot.org/DP00086>

[illegible]

Bioschemas

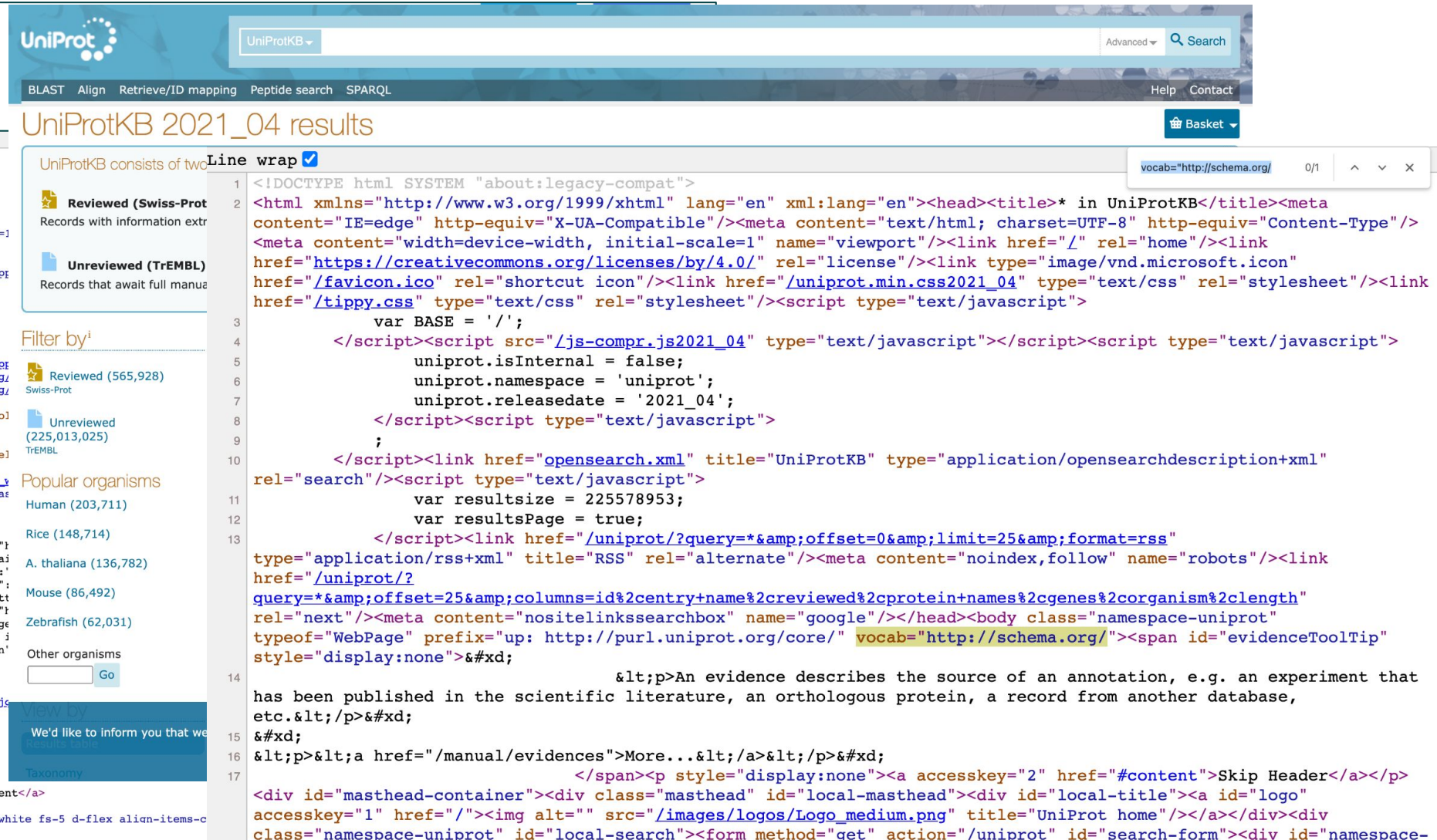


```

1 <!DOCTYPE html>
2 <html lang="en">
3 <html>
4 <head>
5   <meta charset="utf-8">
6   <meta http-equiv="X-UA-Compatible" content="chrome=1">
7   <meta name="viewport" content="width=device-width, initial-scale=1">
8   <meta name="msapplication-TileColor" content="#da532c">
9   <meta name="theme-color" content="#ffffff">
10  <meta property="og:image" content="//bioschemas.org/assets/img/app/
11  </title>
12    Bioschemas - Bioschemas, what and why?
13
14  </title>
15  <meta name="description" content="1. What is Bioschemas?">
16  <!-- Icons -->
17  <link rel="apple-touch-icon" sizes="180x180" href="/assets/img/app/
18  <link rel="icon" type="image/png" sizes="32x32" href="/assets/img/
19  <link rel="icon" type="image/png" sizes="16x16" href="/assets/img/
20  <link rel="manifest" href="/assets/img/site.webmanifest">
21  <link rel="mask-icon" href="/assets/img/safari-pinned-tab.svg" col
22  <link rel="shortcut icon" href="">
23  <!-- Styling -->
24  <link href="https://fonts.googleapis.com/css?family=Noto+Sans" rel
25  <link rel="stylesheet" href="/assets/css/main.css">
26  <link rel="canonical" href="http://bioschemas.org/tutorials/what_y
27  <link rel="alternate" type="application/rss+xml" title="Bioschemas
28
29  <!-- Bioschemas -->
30
31  <script type="application/ld+json">
32    {
33      "@context": "https://schema.org/",
34      "@type": "LearningResource",
35      "@type": "CreativeWork",
36      "@id": "https://bioschemas.org/profiles/Trai
37      [{"@id": "http://edamontology.org/topic_0089"}],
38      "audience": [{"@type": "
39      "Bioschemas"}],
40      "name": "Bioschemas, what and why?",
41      "author": [{"@type": "
42      "Garcia",
43      "@id": "https://bioschemas.org/people/LeylaGarcia",
44      "url": "htt
45      "Arcila",
46      "@id": "https://bioschemas.org/people/RicardoArcila",
47      "url": "htt
48      "Angel",
49      "@id": "https://bioschemas.org/people/VictoriaDominguezDelAngel
50      17",
51      "description": "In this tutorial you will learn what Bioschemas i
52      markup, structured data, bioschemas",
53      "license": "CC-BY 4.0",
54      "version":
55    }
56  </script>
57
58  <!-- JavaScript -->
59  <script src="https://ajax.googleapis.com/ajax/libs/jquery/3.6.0/jq
60  <script src="/assets/js/bootstrap.bundle.min.js"></script>
61  <script src="/assets/js/toc.js"></script>
62  <script defer src="/assets/js/fontawesome-all.js"></script>
63 </head>
64
65 <body>
66   <a class="visually-hidden-focusable" href="#main">Skip to conten
67 <header id="header_wrap" class="p-3">
68   <a class="btn btn-github rounded-bottom rounded-0 rounded-bottom shadow text-d

```

view-source:<https://www.uniprot.org/uniprot>



Custom Types: Bioschemas Proposals

<https://cells.ebisc.org/BIHi006-D> (bioschemas:BioSample)

The screenshot shows the EBISC2 website. At the top is the EBISC2 logo. Below it is a navigation bar with 'Catalogue', 'About', and 'Information for Customers'. A search bar is present with the text 'Search in EBISC cell line'. Below the search bar is a blue banner with the text 'Search for cell line names, genes, cell types, diseases, Biosample-IDs'. To the right of the banner is a button 'Deposit cell line to EBISC'. Below the banner is a red circle. At the bottom left, the text 'BIHi006-D' is displayed in large red font, with 'iPSC line' below it. To the right of this text is a button 'Purchase BIHi006-D' and a link 'Re-stocking'. Below these is the text 'At European Collection of Authenticated Cultures (ECACC)'.

The screenshot shows the Schema.org website. At the top is the Schema.org logo and navigation links 'Documentation', 'Schemas', and 'About'. Below the navigation bar is a search bar. The URL 'https://cells.ebisc.org/BIHi006-D' is entered in the search bar. Below the search bar is a code editor showing the HTML schema. To the right of the code editor is a table titled 'BioSample / Product' with the following columns: 'Property', 'Value', and 'Error'. The table contains the following rows:

Property	Value	Error
@type	BioSample (The type <i>BioSample</i> is not a type defined by the recognised schema (e.g. schema.org).)	4 ERRORS 6 WARNINGS
@id	https://cells.ebisc.org/BIHi006-D	
http://purl.org/dc/terms/conformsTo	https://bioschemas.org/profiles/BioSample/0.1-DRAFT2019_11_12/	
identifier	BIHi006-D	
description	An induced pluripotent stem cell (iPSC) line provided with manually validated scientific and ethical information.	
name	BIHi006-D	
url	https://cells.ebisc.org/BIHi006-D	
url	https://cells.ebisc.org/BIHi006-D	
locationCreated	Germany (The property <i>locationCreated</i> is not recognised by the schema (e.g. schema.org) for an object of type <i>Product</i> .)	
custodian	Berlin Institute of Health (The property <i>custodian</i> is not recognised by the schema (e.g. schema.org) for an object of type <i>Product</i> .)	

<https://validator.schema.org/#url=https://cells.ebisc.org/BIHi006-D>



Non-Conformant Markup

Syntax

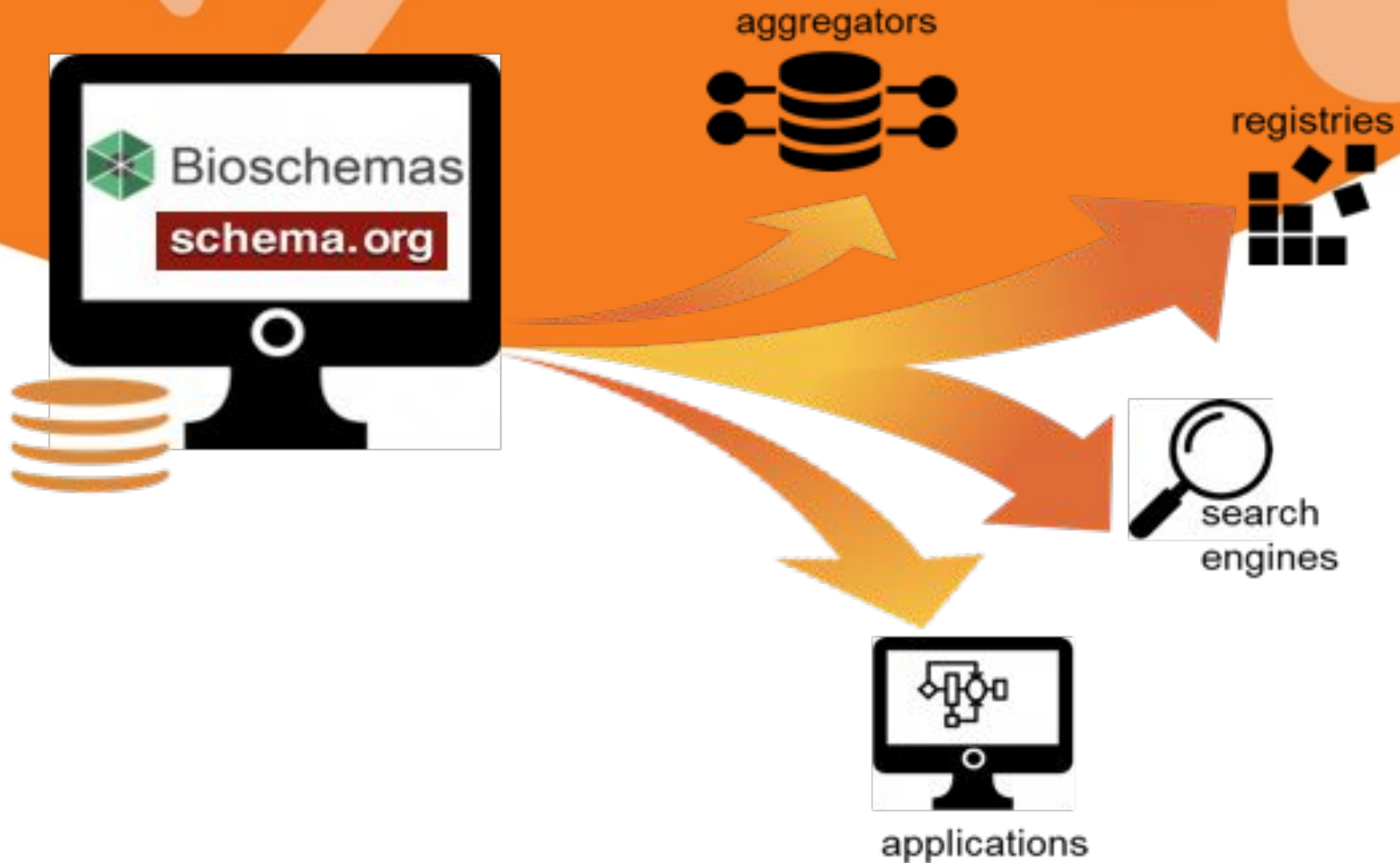
- Invalid (JSON-LD) characters
Very common in database generated pages
- Non-escaped characters, e.g. " included in text
Very common in database generated pages
- Missing commas
- Missing brackets
- Included after </html>
- ...

Semantic/Modelling

- Custom context
- Missing @id
- Missing properties from profile



Markup Harvesting Tools



Markup Harvesting Tools

BMUSE

- Markup extractor
- Markup agnostic
- Adds provenance
- Introduces missing @id
- Single page or whole site
- Web service or standalone app

Gleaner

- Markup extractor
- Validates against Schema.org
- Part of GleanerIO providing triplestore and GraphQL

	<u>BMUSE</u>	<u>Gleaner</u>
Seed	sitemaps, URLs	sitemaps
Static Sites	Yes	Yes
Dynamic Sites	Yes	Yes
JSON-LD	Yes	Yes
RDFa	Yes	No
Bioschemas types	Yes	No
Syntax correction	No	No
Modelling correction	Some	No



BMUSE: Bioschemas Markup Scraper and Extractor

- Data harvester
 - List of URLs and sitemaps
- Extracts markup
 - JSON-LD or RDFa or both
 - Static or dynamic
- Useful for testing deployments
- Building block for
 - Registry population
 - Validation
- Includes provenance of extraction

<https://app.swaggerhub.com/apis-docs/swel/BMUSE/>



BMUSE

0.4.0 OAS3

A Bioschemas project by SWeL funded through Elixir websites so that they are indexable by search engines

Contact the developer

Apache 2.0

Servers

<https://swel.macs.hw.ac.uk/scraper> - Production

developers

GET /getRDF retrieves HTML source from

retrieves HTML source from a given URL and return

Parameters

Name	Description
url required	URL to retrieve source from (must include
string	
(query)	https://ega-archive.org/



Execute

Clear

Responses

Snippets

cURL (bash)

cURL (PowerShell)

cURL (CMD)

```
curl -X 'GET' \
  'https://swel.macs.hw.ac.uk/scraper/getRDF?url=https%3A%2F%2Fega-archive.org%2F&output=jsonld' \
  -H 'accept: */*'
```

Request URL

<https://swel.macs.hw.ac.uk/scraper/getRDF?url=https%3A%2F%2Fega-archive.org%2F&output=jsonld>

Server response

Code Details

200

Response body

```
{
  "result": "success",
  "rdf": [
    {
      "@graph": [
        {
          "http://purl.org/pav/retrievedFrom": [
            {
              "@id": "https://ega-archive.org/"
            }
          ],
          "http://purl.org/pav/createdWith": [
            {
              "@value": "https://github.com/HW-SWeL/BMUSE/releases/tag/0.4.0"
            }
          ],
          "@id": "https://bioschemas.org/crawl/v1/ega-archive//20220106/0",
          "http://purl.org/pav/retrievedOn": [
            {
              "@value": "2022-01-06T11:43:26",
              "@type": "http://www.w3.org/2001/XMLSchema#dateTime"
            }
          ]
        }
      ]
    }
  ]
}
```



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

JSON-LD Playground

[JSON-LD](#) [Playground](#) [Documentation](#) [Developers](#) [Specifications](#) [Branding](#)

JSON-LD Playground

Play around with JSON-LD markup by typing out some JSON below and seeing what gets generated from it at the bottom of the page. Pick any of the examples below to get started.

NOTE: The playground uses [jsonld.js](#) which conforms to [JSON-LD 1.1 syntax \(errata\)](#), [API \(errata\)](#), and [framing \(errata\)](#). Also see the classic [JSON-LD 1.0 playground](#) and the [RDF Distiller](#).

Examples: [Person](#) [Event](#) [Place](#) [Product](#) [Recipe](#) [Library](#) [Activity](#)

[Permalink](#) [Gist](#) [Shortcuts](#)

[JSON-LD Input](#) [Options](#)

```
{
  "@graph": [
    {
      "http://purl.org/pav/retrievedFrom": [
        {
          "@id": "https://ega-archive.org/"
        }
      ],
      "http://purl.org/pav/createdWith": [
        {
          "@value": "https://github.com/HW-SWeL/BMUSE/releases/tag/0.4.0"
        }
      ],
      "@id": "https://bioschemas.org/crawl/v1/ega-archive//20220106/0",
      "http://purl.org/pav/retrievedOn": [
        {
          "@value": "2022-01-06T16:34:54",
          "@type": "http://www.w3.org/2001/XMLSchema#dateTime"
        }
      ]
    }
  ],
}
```

[Expanded](#) [Compacted](#) [Flattened](#) [Framed](#) [N-Quads](#) [Normalized](#) [Table](#)

```
},
{
  "@id": "https://bioschemas.org/crawl/v1/ega-archive//20220106/0/ega-archive.org",
  "@type": [
    "https://schema.org/Organization"
  ],
  "https://schema.org/name": [
    {
      "@value": "Centre for Genomic Regulation (CRG) "
    }
  ],
  "https://schema.org/url": [
    {
      "@id": "https://ega-archive.org/"
    }
  ]
}
```

```
"http://purl.org/pav/createdWith": [
  {
    "@value": "https://github.com/HW-SWeL/BMUSE/releases/tag/0.4.0"
  }
],
"@id": "https://bioschemas.org/crawl/v1/ega-archive//20220106/0",
"http://purl.org/pav/retrievedOn": [
  {
    "@value": "2022-01-06T16:34:54",
    "@type": "http://www.w3.org/2001/XMLSchema#dateTime"
  }
],
}
```

[Expanded](#) [Compacted](#) [Flattened](#) [Framed](#) [N-Quads](#) [Normalized](#) [Table](#) [Visualized](#) [Signed with RSA](#) [Signed](#)



<https://json-ld.org/playground/>

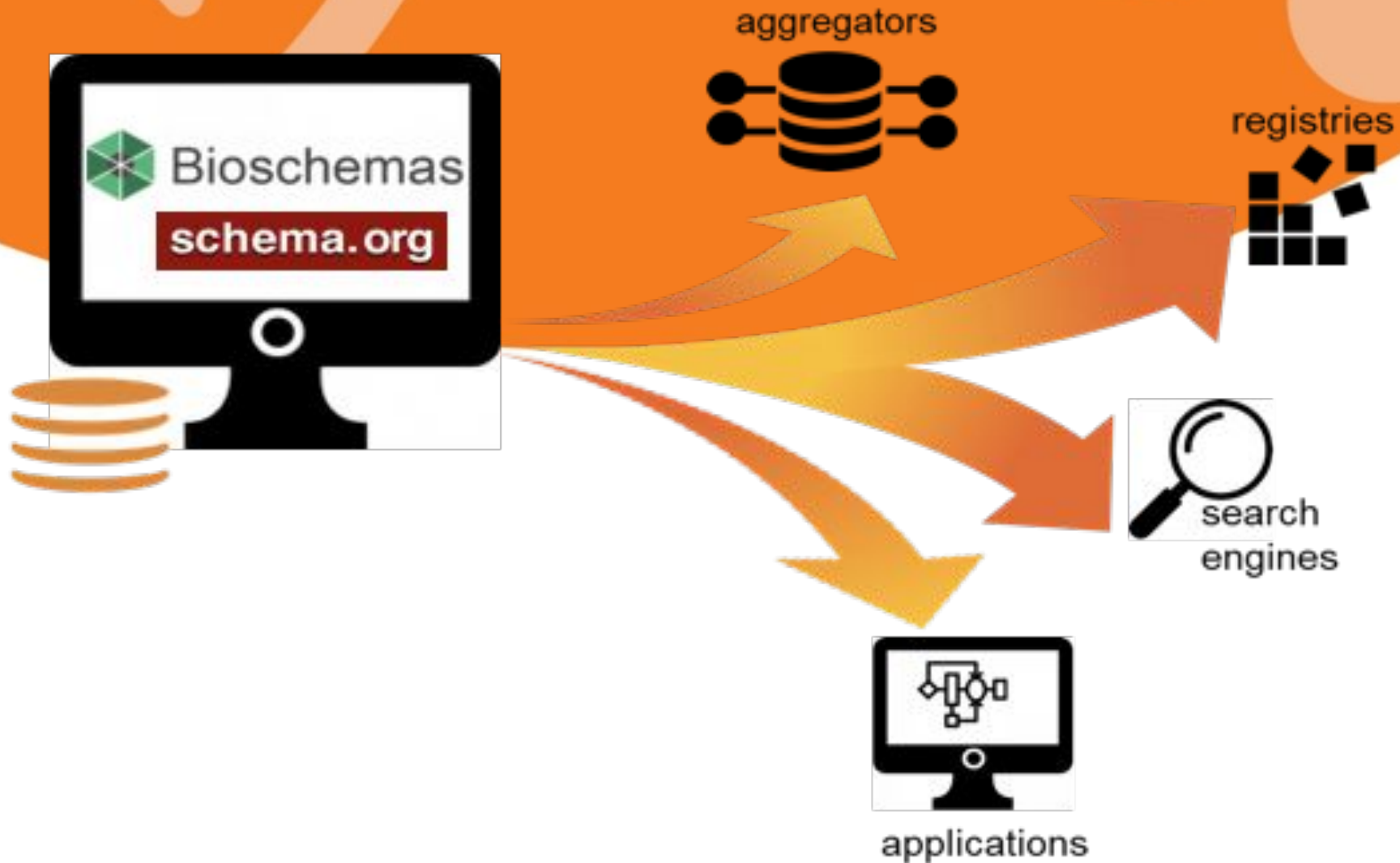
Hands-on: Retrieving and Exploring Markup

- Use BMUSE to retrieve a page
<https://app.swaggerhub.com/apis-docs/swel/BMUSE/>
- Explore the retrieved markup using JSON-LD Playground
<https://json-ld.org/playground/>

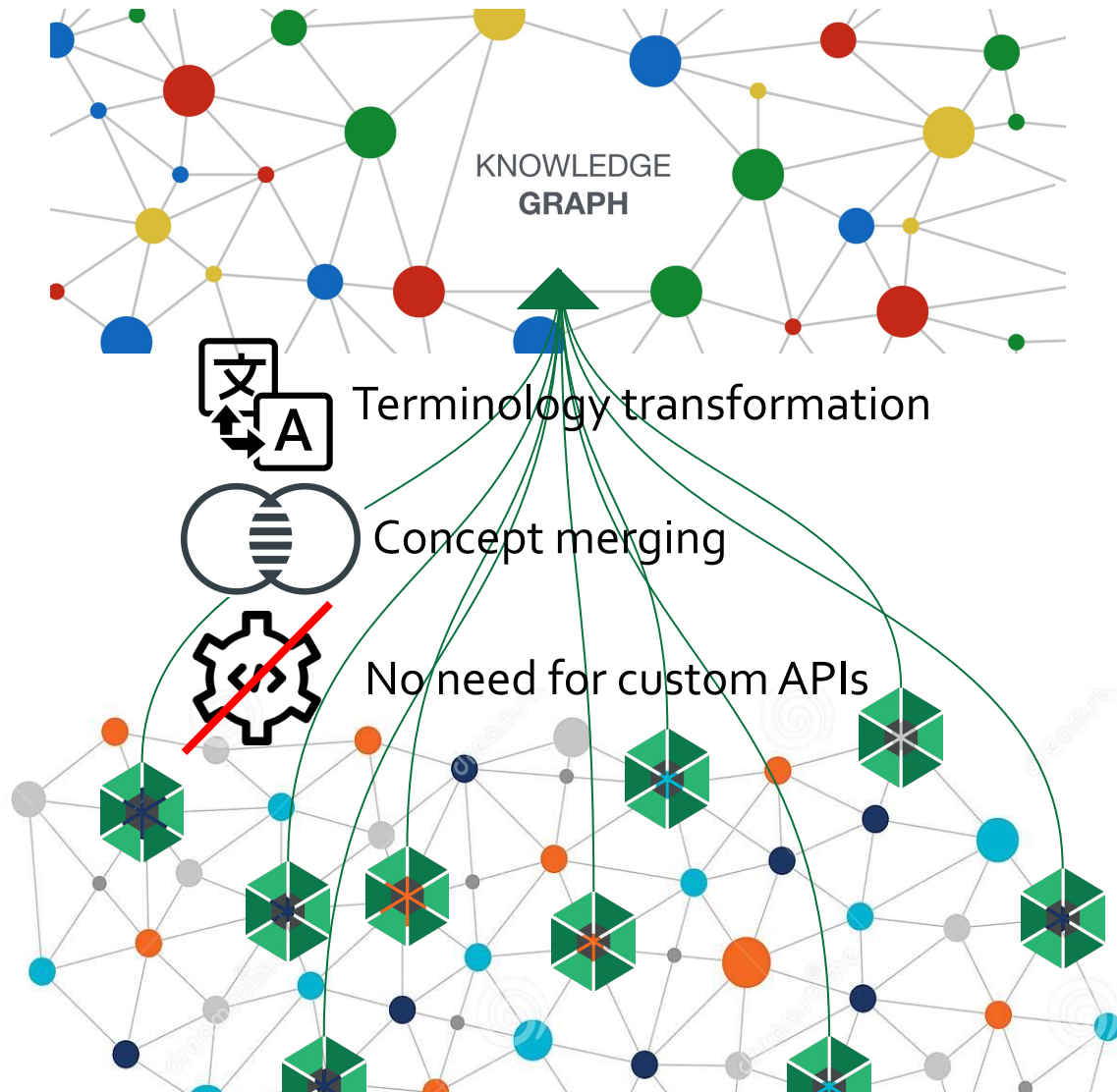
The image displays two web interfaces side-by-side. The left interface is the SwaggerHub page for the BMUSE API. It features a dark header with the SwaggerHub logo and 'Sign Up'/'Log In' buttons. The main content area shows the API title 'BMUSE' with version '0.4.0' and 'OAS3' specification. A description states it's a Bioschemas project by SWEL funded through Elixir-Excellerate, aimed at improving data findability by scraping JSON-LD or RDFa from URLs and returning triples in a JSON wrapper. It also mentions 'Contact the developer' and 'Apache 2.0' license. Below this, a 'Servers' section lists 'https://swel.macs.hw.ac.uk/scrapper - Production sv'. A 'developers' section shows a 'GET /getRDF' endpoint that 'retrieves HTML source from a given URL and returns the HTML source'. A 'Parameters' table lists a required 'url' parameter of type 'string' with a 'query' parameter, with an example value 'https://ega-archive.org/'. The right interface is the JSON-LD Playground. It has a top navigation bar with links for 'JSON-LD', 'Playground', 'Documentation', 'Developers', 'Specifications', and 'Branding'. The main heading is 'JSON-LD Playground'. A note explains that the playground uses 'jsonld.js' conforming to JSON-LD 1.1 syntax, API, and framing, and also mentions the classic JSON-LD 1.0 parser. Below the note are 'Examples' for Person, Event, Place, Product, Recipe, Library, and Activity, with a 'Permalink' button. The 'JSON-LD Input' section shows a JSON-LD document with a graph containing information about a retrieved page from 'https://ega-archive.org/' and its creation with a specific GitHub release. The 'Options' section is visible. At the bottom, there are buttons for 'Expanded', 'Compacted', 'Flattened', 'Framed', 'N-Quads', 'Normalized', 'Table', 'Visualized', and 'Signed with RSA'. The 'Expanded' view shows the resulting JSON-LD document with expanded URIs and types.



Markup to Knowledge Graph

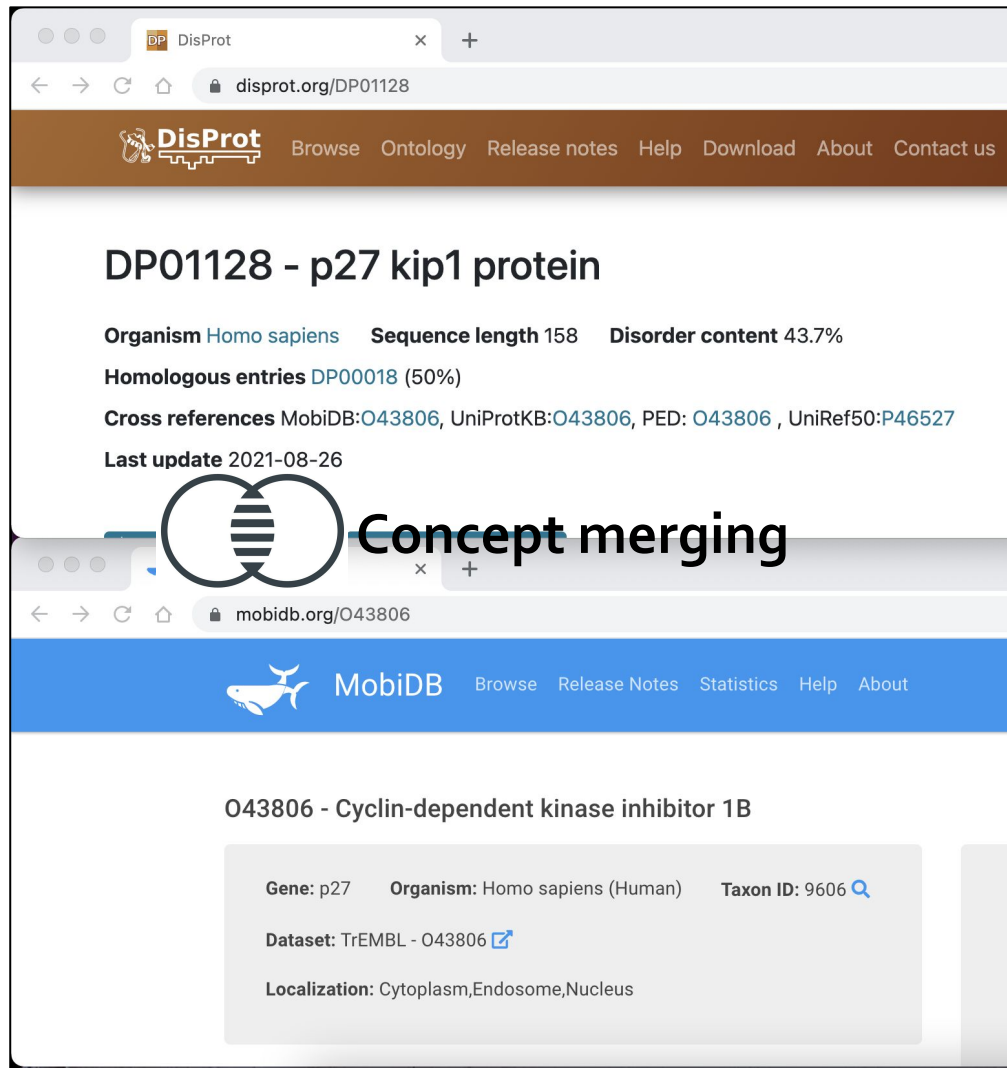


Harvested Markup to Knowledge Graph



- Markup harvested through standard API
 - HTTP Get requests
 - Saves time as common API and model (Bioschemas) for all sites
- Harvested markup is page centric
 - Multiple sites represent the same concept
 - Sites use their own IDs
- Needs to be concept centric
 - Reconcile IDs
 - Merge entries
- Model transformation

Merging Entries Options



The image shows two overlapping browser windows. The top window is DisProt (disprot.org/DP01128) for the entry 'DP01128 - p27 kip1 protein'. It lists organism as Homo sapiens, sequence length as 158, and disorder content as 43.7%. It also shows homologous entries and cross references. The bottom window is MobiDB (mobidb.org/O43806) for the entry 'O43806 - Cyclin-dependent kinase inhibitor 1B'. It lists gene as p27, organism as Homo sapiens (Human), taxon ID as 9606, dataset as TrEMBL - O43806, and localization as Cytoplasm, Endosome, Nucleus. A large circular arrow icon with the text 'Concept merging' is overlaid on the intersection of the two windows, indicating the process of combining information from different sources.

1. Pick canonical source

Useful for values like name

- Supplement with other entries when values are missing

2. Concatenate entries from all sources

Useful for values like synonym

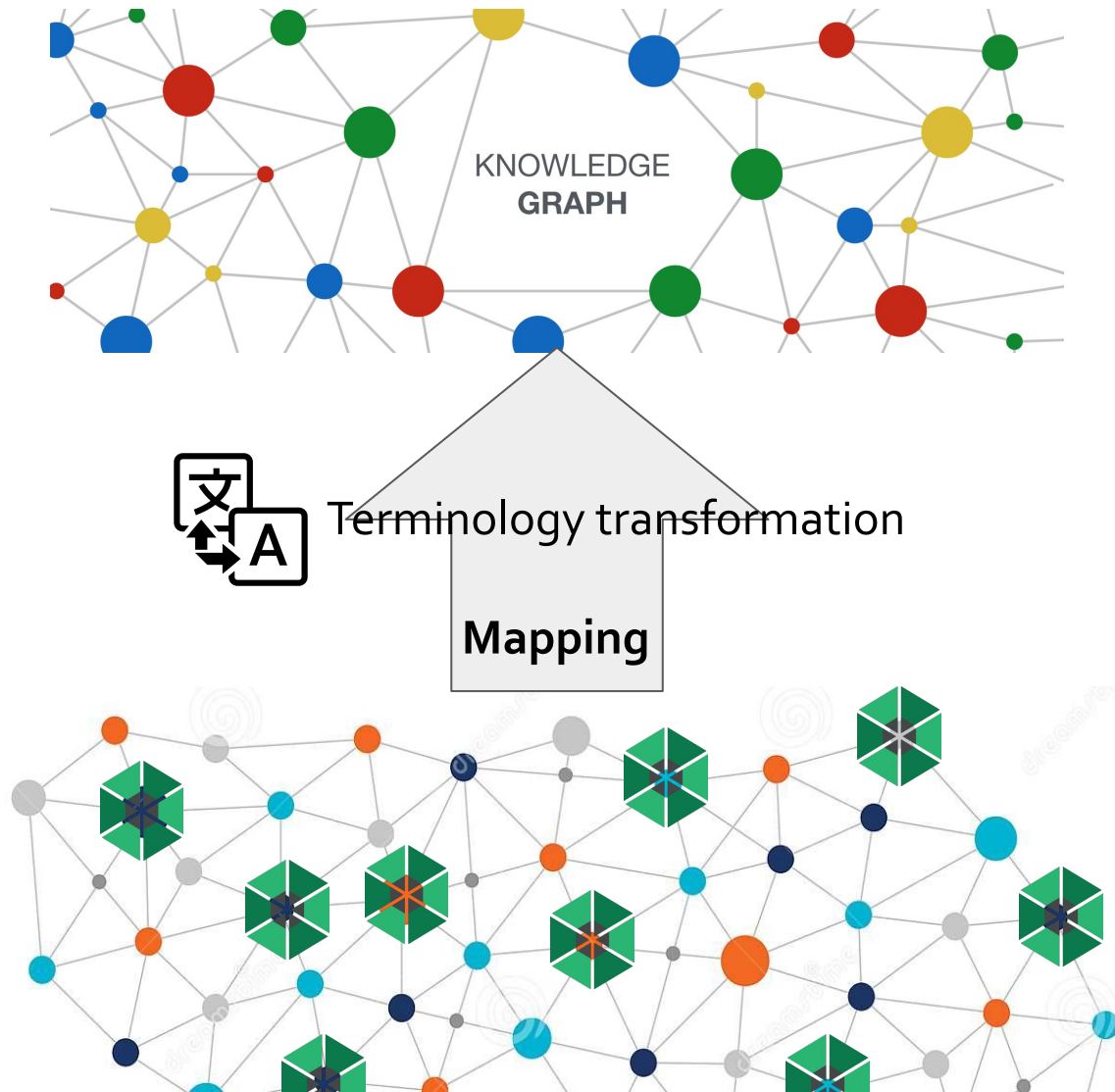
3. Keep all values with provenance of source

Wikidata approach

Choice will depend on application needs!

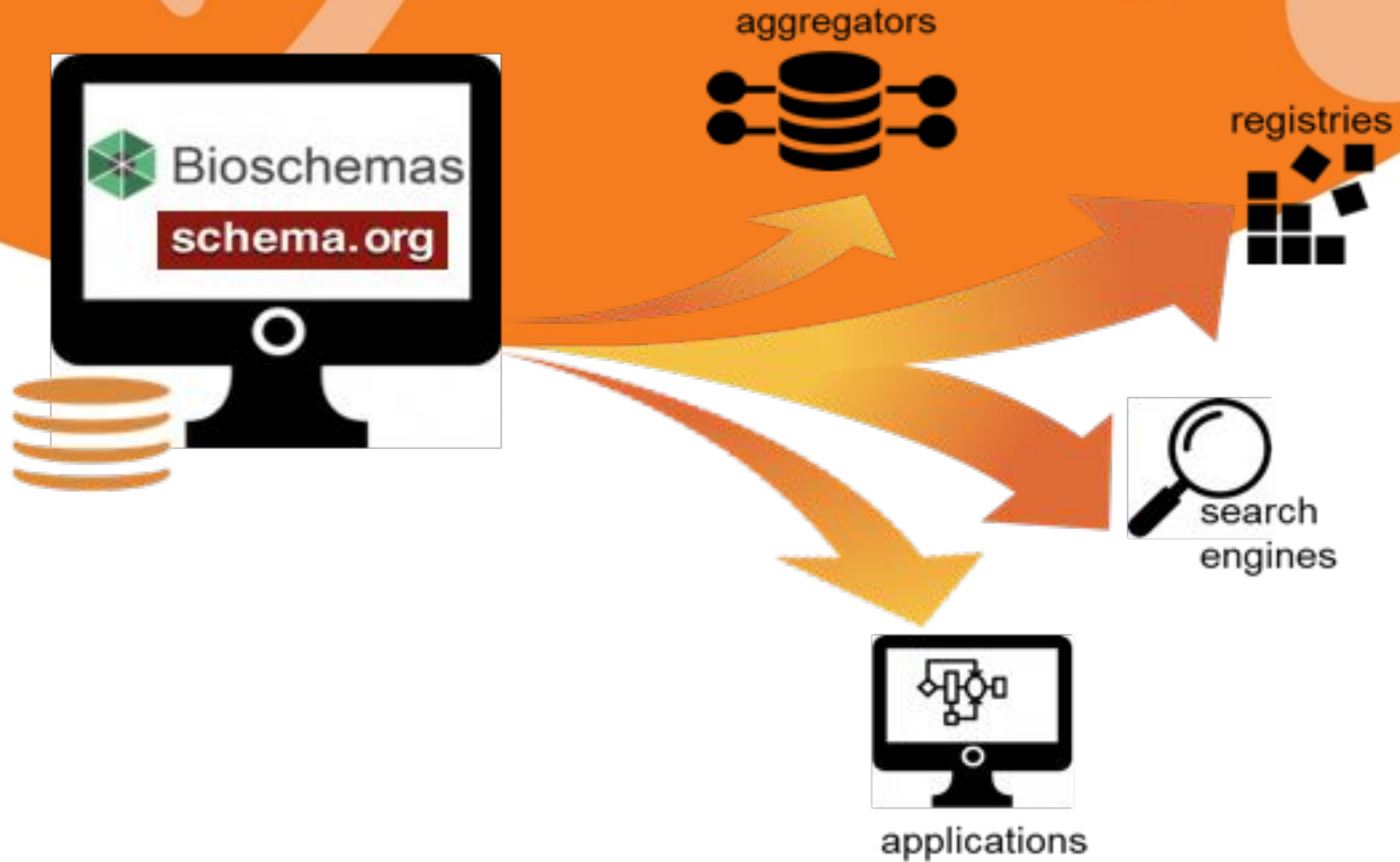


Model Transformation



1. Identify common concepts and properties between models
2. Create mapping between concepts and properties
3. Apply mapping to harvested data

Existing Markup



Bioschemas Live Deployment List

bioschemas.org/liveDeploys



[Home](#) [Getting Started](#) [Specifications](#) [Deploy & Develop](#) [Community](#) [About](#)

Sites Implementing Bioschemas's Markup

- 77 sites with Bioschemas markup
- 155 profile deployments

By Site By Profile By ELIXIR Node

Profiles: 2	AIP Mutation Database	▼
Profiles: 2	Alliance of Genome Resources	▼
Profiles: 1	BAR 3.0	▼
Profiles: 3	Bgee	▼
Profiles: 2	bio.tools	▲
- Sitemap: https://bio.tools/api/sitemap.xml - Nodes: DK, FR, ES Implemented Profiles ComputationalTool v1.0-RELEASE 22,687 tools FormalParameter v1.0-RELEASE Page SMV BMUSE Page SMV BMUSE		
Profiles: 1	Biological Magnetic Resonance Data Bank (BMRB)	▼
Profiles: 2	Biosamples	▼
Profiles: 1	Bioschemas	▼

bioschemas.github.io/live_depi x +
github.com/BioSchemas/bioschemas.github.io/blob/master/_data/live_deployments.json

1465 lines (1465 sloc) 44.2 KB

```
1 {
2   "@context": {
3     "conformsTo": "http://purl.org/dc/terms/conformsTo",
4     "description": "https://schema.org/description",
5     "exampleURL": "https://schema.org/url",
6     "highlight": "http://example.com/bs/highlight",
7     "keywords": "https://schema.org/keywords",
8     "name": "https://schema.org/name",
9     "nodes": "http://example.com/bs/nodes",
10    "profiles": "http://example.com/bs/profiles",
11    "profileName": "https://schema.org/name",
12    "resources": "https://schema.org/Website",
13    "sitemap": "http://example.com/bs/sitemap",
14    "url": "https://schema.org/url"
15  },
16  "resources": [
17    {
18      "name": "Identifiers.org",
19      "keywords": ["RIR"],
20      "url": "https://identifiers.org/",
21      "nodes": "EBI",
22      "profiles": [
23        {
24          "profileName": "DataCatalog",
25          "conformsTo": "0.1",
26          "exampleURL": "https://identifiers.org/"
27        }
28      ]
29    },
30    {
31      "name": "FAIRsharing",
32      "keywords": ["RIR"],
33      "url": "https://fairsharing.org/",
34      "nodes": "UK",
35      "sitemap": "https://fairsharing.org/sitemap.xml",
36      "profiles": [
37        {
38          "profileName": "DataCatalog",
39          "conformsTo": "0.1"
```


Project 29: Bioschemas

Local



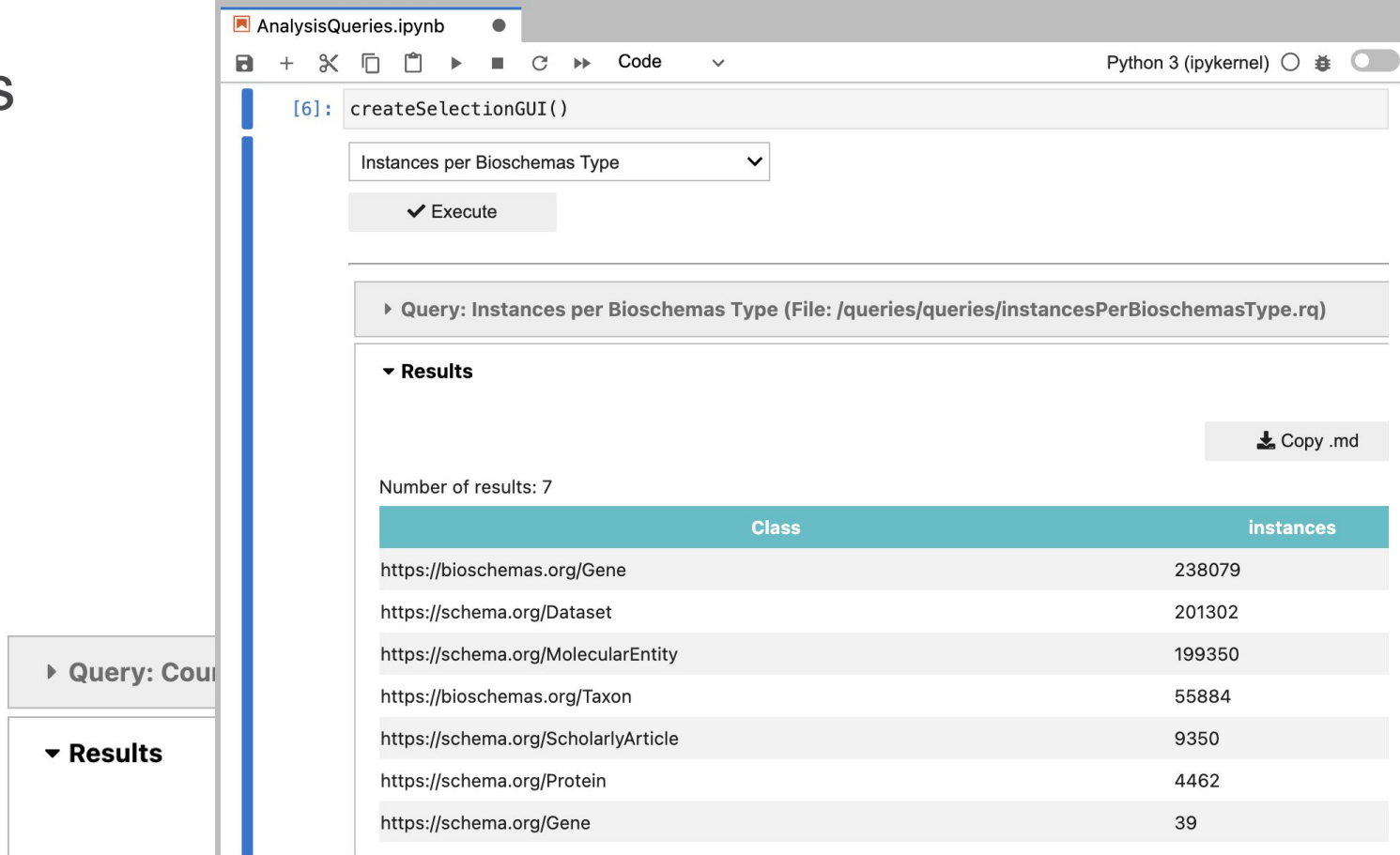
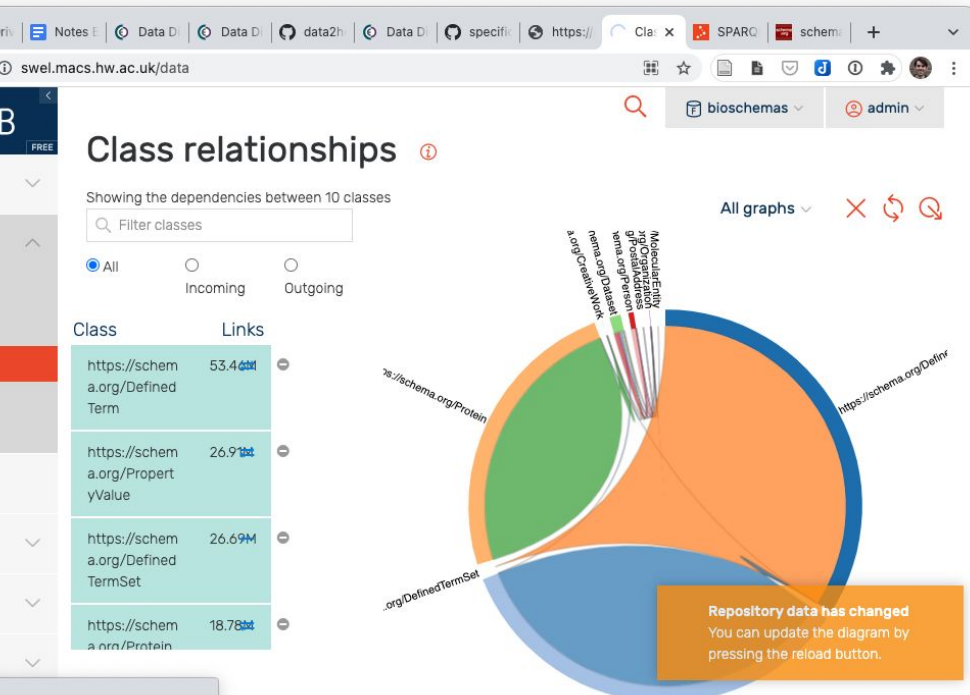
bioschemas.org

total statements
11,839,960

Sites harvested
25

Pages harvested
413,748

<https://swel.macs.hw.ac.uk/data/>



Number of results: 146

domain	type	count
www.gbif.org	https://schema.org/PostalAddress	297090
www.gbif.org	https://schema.org/Person	291260
bgee.org	https://bioschemas.org/Gene	263793
www.gbif.org	https://schema.org/Organization	186688
www.gbif.org	https://schema.org/PropertyValue	126268
massbank.eu	https://schema.org/Dataset	76249
massbank.eu	https://schema.org/MolecularEntity	76249

Exploring Bioschemas Harvested Data



AnalysisQueries.ipynb

File Edit View Run Kernel Tabs Settings Help

ANALYSISQUERIES.IPYNB

<> M

1. Analysis of Harvested Bioschemas Data

1.1. Introduction

1.1.1. HCLS Dataset Statistics

1.1.2. Bioschemas Analysis Queries

1.2. Query Execution

1.2.1. Library Imports

1.2.2. Result Display Function

1.2.3. Loading and Querying IDP-KG

1.3. Query Interface

1. Analysis of Harvested Bioschemas Data

Authors:
Alasdair J G Gray (ORCID:0000-0002-5711-4872), Heriot-Watt University, Edinburgh, UK
Imran Asif (ORCID:0000-0002-1144-6265), Heriot-Watt University, Edinburgh, UK
Alban Gaignard

License: Apache 2.0

Acknowledgements: This work was funded as part of the [ELIXIR Infrastructure](#) project. This notebook has been adapted from the [Knowledge Graph \(IDP-KG\)](#).

1.1. Introduction

In this notebook we supply a user interface to query the harvested Bioschemas data repository. Details of the data harvested and loaded into the triplestore are available in the [HCLS Dataset Description Community Profile](#).

The available queries can be grouped as discussed in the following summary:

1.1.1. HCLS Dataset Statistics

As part of the W3C Health Care and Life Sciences Interest Group [guidelines](#), we provide useful statistics to understand a dataset. The queries range from providing simple statistics such as the number of typed objects linked to more complex statistics such as the number of typed objects linked to a specific domain.

1. Number of triples
2. Number of unique typed entities
3. Number of unique subjects
4. Number of unique properties
5. Number of unique objects
6. Number of unique classes

AnalysisQueries.ipynb

File Edit View Run Kernel Tabs Settings Help

ANALYSISQUERIES.IPYNB

<> M

1. Analysis of Harvested Bioschemas Data

1.1. Introduction

1.1.1. HCLS Dataset Statistics

1.1.2. Bioschemas Analysis Queries

1.2. Query Execution

1.2.1. Library Imports

1.2.2. Result Display Function

1.2.3. Loading and Querying IDP-KG

1.3. Query Interface

1.3. Query Interface

The following cell generates the UI and allows the selection of query to run.

The results of the query execution are displayed below. The query box can be expanded to show the text of the query. The results box is scrollable.

Descriptions of the first 14 queries can be found in the [HCLS Dataset Description Community Profile](#). The other queries are explained [above](#).

```
[7]: createSelectionGUI()
```

Count of Types per Domain

Count of Types per Domain
This query finds the number of instances of types per web domain

PREFIX schema: <https://schema.org/>
PREFIX pav: <http://purl.org/pav/>
SELECT ?domain ?type (COUNT(*) AS ?count)
WHERE {
 GRAPH ?g {
 ?s a ?type
 }
}

Execute

Query: (File: queries/countTypeInstancePerDomain.rq)

Results

Copy .md

Number of results: 161

domain	type	count
bgee.org	https://bioschemas.org/Gene	263,793
bgee.org	https://bioschemas.org/Taxon	49,088
bgee.org	https://schema.org/WebPage	49,009

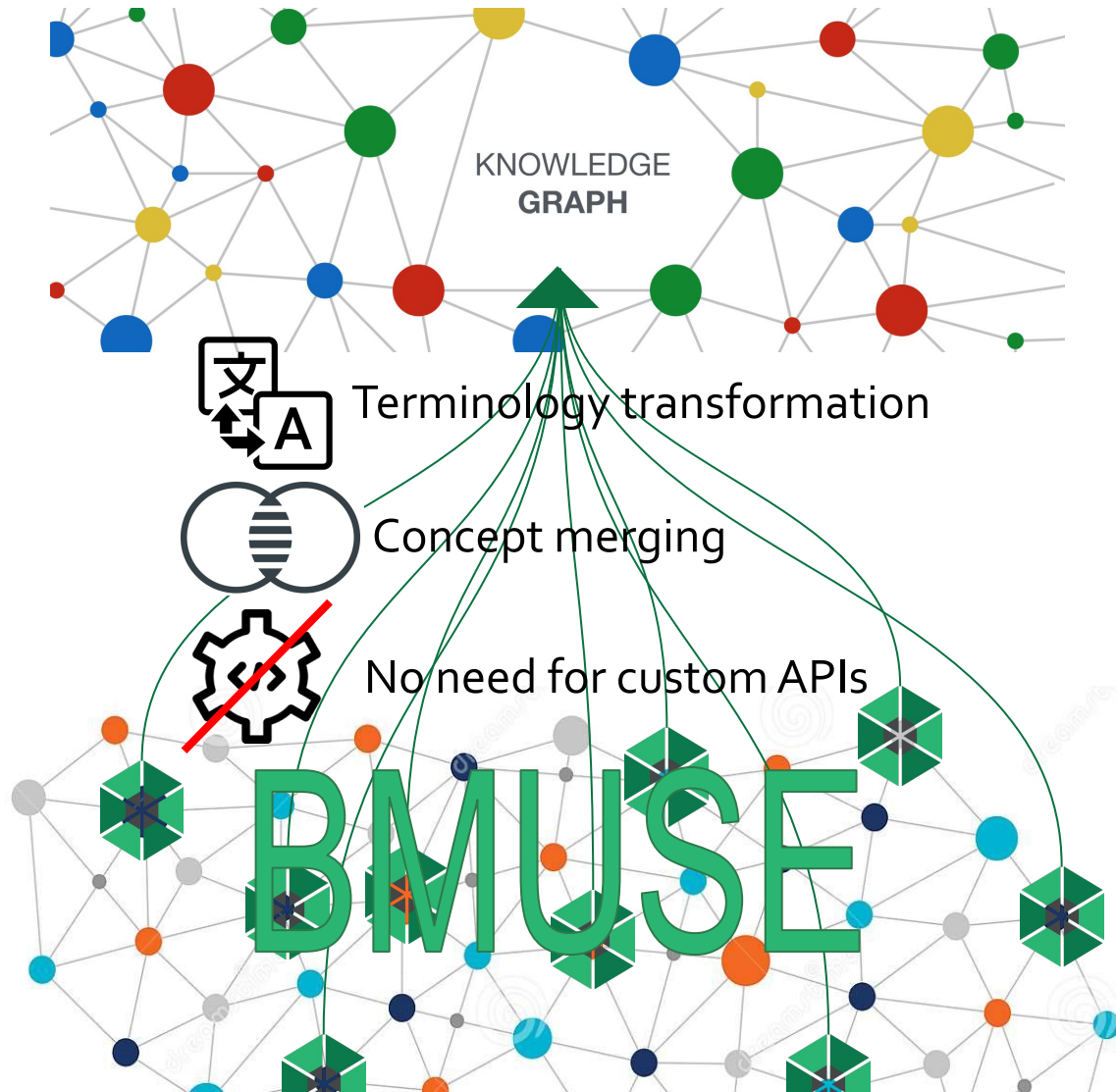
Simple 0 1 Python 3 (ipykernel) | Idle Mem: 146.91 / 2048.00 MB

Simple 0 1 Python 3 (ipykernel) | Idle Mem: 154.94 / 2048.00 MB

Mode: Command Ln 1, Col 1 AnalysisQueries.ipynb



Summary



- Markup harvested through standard API
- Markup is often messy and inaccurate
- Need a data harvester tool, e.g. BMUSE
- Markup is page centric, common concepts need to be merged
- Over 150 deployments to be exploited!
 - Bioschemas has harvested a small fraction