BioHackrXiv

Bioschemas data harvesting project report

Alasdair Gray[1], Petros Papadopoulos[1], Alban Gaignard[2], Thomas Rosnet[2], Ivan Mičetić[3], and Sébastien Moretti[4]

1 Heriot-Watt University, Edinburgh, UK **2** Nantes Université, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France **3** University of Padua, Padova, Italy **4** Université de Lausanne, Lausanne, Switzerland

BioHackathon series:
BioHackathon Europe 2021
Barcelona, Spain, 2021
Bioschemas team

Submitted: 23 Mar 2022

Published by BioHackrXiv.org

Introduction

The promise of Bioschemas is that it makes consuming data from multiple resources more straightforward. However, this hypothesis has not been tested by conducting a large scale harvest of deployed markup and making this available for others to reuse. Therefore, the goal of this hackathon project is to harvest a collection of Bioschemas markup from a number of different sites listed on the Bioschemas live deploys page using the Bioschemas Markup Scraper and Extractor (BMUSE). The harvested data will be made available for others and loaded into a triplestore to allow for further exploration.

Data Harvesting

Prior to the BioHackathon, we set about harvesting data from as many of the Bioschemas live deploy sites as possible. At the time of the BioHackathon, there were 70 sites listed, and 137 profile deployments (a site can deploy multiple profiles, e.g. Dataset and DataCatalog). Not all deployments could be harvested since they do not provide sitemaps listing the pages within the site. At the time of the BioHackathon there were 25 sites with sitemaps. Several of these do not list the pages containing data, limiting the amount that could be harvested.

The list of sites to be harvested were gathered in a GitHub project board so that progress could be tracked. The cards in this board were annotated to state whether the source was known to use a static site deployment (i.e. the markup is embedded in the page source by the server) or dynamic single page application (i.e. the page content is generated client side using Javascript), and also whether they were known to have data content or limited content of Dataset and DataCatalog.

The Bioschemas Markup Scraper and Extractor (BMUSE) was used for the harvesting of the data. During the harvesting we found a two key issues with BMUSE which arose due to the scale of the data harvest. The first was that errors in the JSON-LD were not correctly identified and logged. The second was a memory limit relating to JSoup which meant that only about 24,104 pages were scraped out of the 50,000 in the sitemap file (BMUSE #82). Fixes to these issues were applied resulting in BMUSE v0.5.2 being used for most of the harvesting.

The data harvesting workflow consisted of the following steps:

1. Pick one of the sites to be harvested: priority was given to static sites with data content since these could be harvested more quickly and went beyond Dataset/DataCatalog markup.
2. For each sitemap in the sitemap index, harvest the content from the source pages.
3. Merge the individual nquad files for each page in the (sub)sitemap into a single nquad file.
4. Load the merged nquad file into the triplestore.
5. Make the merged nquad file available on the web.
6. Update the project README with details of what had been harvested.

Where issues were found with the source site, these were fedback to the data provider to allow them to revise their markup. For example, it was found that MassBank were including characters in their string values such as " that need to be escaped to generate valid JSON-LD (MassBank-Web #316).

In total, six sites were found to be unscrapable. These were

- InterPro: a dynamic site providing a sitemap. However, the sitemap did not conform with the sitemap standard.
- Scholia COVID-19 URL list: a site generated via SPARQL queries over the Wikidata endpoint. Unable to scrape due to timeout being reached before the data being available.
- SwissModel: the provided sitemap did not conform with the sitemap stanard.
- WikiPathways: sitemap was empty.
- IPPIDB: a dynamic site with data content corresponding to MolecularEntities. However, the pages exhibiht inconsistent rendering when tested in the browser and could not be harvested with BMUSE.
- OrphaNet: a static site with disease markup. The sitemap conforms with the older Google proposal for sitemaps rather than the widely used 0.9 version expected by BMUSE

Data Analysis

We reused the notebook originally developed at BioHackathon 2020 (Gray et al., 2021) and since evolved for the Intrinsically Disordered Protein Knowledge Graph (IDP-KG) (Gray et al., 2022). We include the HCLS Dataset Description profile statistics queries[1] (Dumontier et al., 2016), read in from an existing repository. We also include queries developed specifically for the analysis of the Bioschemas harvested data.

To use the notebook (MyBinder launcher), you simply need to run all cells and then select the query you would like to execute from the resulting dropdown menu.

We now present the results of the queries obtained during the hackathon, i.e. the data values are as they were on 11 November 2021. Running the notebook in March 2022 obtains different results due to more harvested data having been added.

HCLS Dataset Statistics

We include here a selection of results from some of the HCLS statistics queries. We focus on those providing the most interesting statistics for the available data. For the full set of queries and results, please run the notebook.

Number of triples

This is the raw count of the number of triples contained in the triplestore repository.

triples
10,610,743

Number of named graphs

The result presented here is equivalent to number of pages harvested since BMUSE generates a named graph for each page harvested.

[1] Dataset Descriptions: HCLS Community Profile §6 accessed March 2022

graphs
413,748

Number of instance per class

There are many different types included in the markup. BMUSE extracts all markup, not just Bioschemas profiles.

The results are ordered by the Class IRI; in the notebook you can edit the query and change the ordering of results.

(57 results)

Class	distinctInstances
http://rdfs.org/sioc/ns#Item	57
http://xmlns.com/foaf/0.1/Document	89
http://xmlns.com/foaf/0.1/Image	219
https://bioschemas.org/Gene	238,079
https://bioschemas.org/Protein	1,262
https://bioschemas.org/Taxon	55,884
https://schema.org/AboutPage	1
https://schema.org/Action	3
https://schema.org/Answer	8
https://schema.org/BioChemEntity	49,823
https://schema.org/BreadcrumbList	14,037
https://schema.org/ChemicalSubstance	29
https://schema.org/CollectionPage	187
https://schema.org/CollegeOrUniversity	2
https://schema.org/ContactPoint	148
https://schema.org/CreativeWork	14,299
https://schema.org/DataCatalog	7,439
https://schema.org/DataDownload	1,497
https://schema.org/Dataset	201,302
https://schema.org/DefinedTerm	4,261
https://schema.org/DefinedTermSet	4,112
https://schema.org/DigitalDocument	1
https://schema.org/EducationalOrganization	3
https://schema.org/Event	12,818
https://schema.org/FAQPage	1
https://schema.org/Gene	39
https://schema.org/GeoShape	19,398
https://schema.org/GovernmentOrganization	1
https://schema.org/ItemList	187
https://schema.org/ListItem	28,137
https://schema.org/MolecularEntity	199,350
https://schema.org/NGO	11,717
https://schema.org/Offer	5
https://schema.org/Organization	206,715
https://schema.org/PeopleAudience	2,475
https://schema.org/Person	326,935
https://schema.org/Place	19,438
https://schema.org/PostalAddress	307,406
https://schema.org/PropertyValue	144,002
https://schema.org/Protein	4,462

Class	distinctInstances
https://schema.org/QAPage	1
https://schema.org/Question	8
https://schema.org/ScholarlyArticle	9,350
https://schema.org/SearchAction	5
https://schema.org/SequenceAnnotation	15,786
https://schema.org/SequenceRange	15,786
https://schema.org/SoftwareApplication	4
https://schema.org/SoftwareSourceCode	4
https://schema.org/Study	4,328
https://schema.org/Thing	27,872
https://schema.org/URL	1
https://schema.org/WebApplication	3
https://schema.org/WebPage	55,114
https://schema.org/WebSite	5
https://schema.org/contact	40
https://schema.org/hostInstitution	40
https://schema.org/url	10,360

Bioschemas Queries

The following queries focus on features of interest to the Bioschemas community.

Instances per Bioschemas Class

Note that due to the data content we need to include some properties with both a Bioschemas namespace and a Schema.org namespace.

The results are ordered by the count of the number of instances; in the notebook you can edit the query and change the ordering of results.

(18 results)

Class	instances
https://schema.org/Person	326,935
https://bioschemas.org/Gene	238,079
https://schema.org/Organization	206,715
https://schema.org/Dataset	201,302
https://schema.org/MolecularEntity	199,350
https://bioschemas.org/Taxon	55,884
https://schema.org/BioChemEntity	49,823
https://schema.org/SequenceAnnotation	15,786
https://schema.org/SequenceRange	15,786
https://schema.org/Event	12,818
https://schema.org/ScholarlyArticle	9,350
https://schema.org/DataCatalog	7,439
https://schema.org/Protein	4,462
https://schema.org/Study	4,328
https://bioschemas.org/Protein	1,262
https://schema.org/Gene	39
https://schema.org/ChemicalSubstance	29
https://schema.org/SoftwareApplication	4

Number of Domains

This result informs us how many web domains were harvested. This is approximately equal to the number of datasets, but some sites may host more than one dataset so not necessarily an exact correspondence.

count
25

Number of Pages per Domain

We now report the number of pages that have been harvested from each domain. Note that we do not understand the empty domain as all markup was extracted from a web domain.

(25 results)

domain	count
massbank.eu	76,253
scholia.toolforge.org	74,319
www.gbif.org	68,167
test.intermine.org	49,959
bgee.org	49,022
www.metanetx.org	49,012
tess.elixir-europe.org	13,939
ega-archive.org	11,833
fairsharing.org	6,351
prosite.expasy.org	5,858
ippidb.pasteur.fr	2,433
mobidb.org	2,082
disprot.org	2,043
pcddb.cryst.bbk.ac.uk	1,402
www.ebi.ac.uk	672
proteinensemble.org	187
www.france-bioinformatique.fr	86
pairedomicsdata.bioinformatics.nl	78
www.covid19dataportal.org	19
	12
www.alliancegenome.org	11
biopragmatics.github.io	3
nanocommons.github.io	3
bridgedb.github.io	2
www.uniprot.org	2

Count of Types per Domain

We now report the number of instances of each type on each domain. What is intersting here is the fact that Bgee has many proteins listed on their pages.

The results are ordered by the count of the number of instances; in the notebook you can edit the query and change the ordering of results.

(146 results)

domain	type	count
www.gbif.org	https://schema.org/PostalAddress	297,090
www.gbif.org	https://schema.org/Person	291,260
bgee.org	https://bioschemas.org/Gene	263,793
www.gbif.org	https://schema.org/Organization	186,688
www.gbif.org	https://schema.org/PropertyValue	126,268
massbank.eu	https://schema.org/Dataset	76,249
massbank.eu	https://schema.org/MolecularEntity	76,249
scholia.toolforge.org	https://schema.org/CreativeWork	74,310
scholia.toolforge.org	https://schema.org/MolecularEntity	74,310
www.gbif.org	https://schema.org/Dataset	63,134
test.intermine.org	https://schema.org/Dataset	49,959
test.intermine.org	https://schema.org/BioChemEntity	49,823
bgee.org	https://bioschemas.org/Taxon	49,059
bgee.org	https://schema.org/WebPage	49,009
www.metanetx.org	https://schema.org/CreativeWork	49,002
www.metanetx.org	https://schema.org/MolecularEntity	49,001
prosite.expasy.org	https://schema.org/Person	31,364
tess.elixir-europe.org	https://schema.org/ListItem	27,872
tess.elixir-europe.org	https://schema.org/Thing	27,872
www.gbif.org	https://schema.org/GeoShape	19,398
www.gbif.org	https://schema.org/Place	19,398
tess.elixir-europe.org	https://schema.org/BreadcrumbList	13,938
tess.elixir-europe.org	https://schema.org/Event	12,778
prosite.expasy.org	https://schema.org/Organization	11,715
prosite.expasy.org	https://schema.org/NGO	11,714
disprot.org	https://schema.org/PropertyValue	11046
disprot.org	https://schema.org/SequenceAnnotation	11,046
disprot.org	https://schema.org/SequenceRange	11,046
prosite.expasy.org	https://schema.org/url	10,360
tess.elixir-europe.org	https://schema.org/PostalAddress	10,316
ega-archive.org	https://schema.org/DataCatalog	7,431
ega-archive.org	https://schema.org/Dataset	7,431
tess.elixir-europe.org	https://schema.org/Organization	7,110
prosite.expasy.org	https://bioschemas.org/Taxon	6,796
prosite.expasy.org	https://schema.org/ScholarlyArticle	6,681
disprot.org	https://schema.org/DefinedTerm	6,599
fairsharing.org	https://schema.org/Dataset	6,328
prosite.expasy.org	https://schema.org/WebPage	6,093
prosite.expasy.org	https://schema.org/CreativeWork	5,857
fairsharing.org	https://schema.org/CreativeWork	5,542
mobidb.org	https://schema.org/PropertyValue	4,486
mobidb.org	https://schema.org/SequenceAnnotation	4,486
mobidb.org	https://schema.org/SequenceRange	4,486
ega-archive.org	https://schema.org/Study	4,328
disprot.org	https://schema.org/DefinedTermSet	4,076
mobidb.org	https://schema.org/DefinedTerm	3,400
mobidb.org	https://schema.org/DefinedTermSet	3,400
tess.elixir-europe.org	https://schema.org/Person	3,298
tess.elixir-europe.org	https://schema.org/CreativeWork	2,876
disprot.org	https://schema.org/ScholarlyArticle	2,857
tess.elixir-europe.org	https://schema.org/PeopleAudience	2,475
proteinensemble.org	https://schema.org/PropertyValue	2,202
mobidb.org	https://schema.org/CreativeWork	2,073

domain	type	count
mobidb.org	https://schema.org/Protein	2,073
disprot.org	https://schema.org/CreativeWork	2,038
disprot.org	https://schema.org/Protein	2,038
proteinensemble.org	https://schema.org/DefinedTerm	1,626
pcddb.cryst.bbk.ac.uk	https://schema.org/Organization	1,402
pcddb.cryst.bbk.ac.uk	https://schema.org/DataDownload	1,394
prosite.expasy.org	https://bioschemas.org/Protein	1,376
pcddb.cryst.bbk.ac.uk	https://schema.org/Dataset	697
pcddb.cryst.bbk.ac.uk	https://schema.org/Person	697
biopragmatics.github.io	https://schema.org/Dataset	287
www.france-bioinformatique.fr	https://schema.org/ListItem	265
proteinensemble.org	https://schema.org/Protein	254
proteinensemble.org	https://schema.org/SequenceAnnotation	254
proteinensemble.org	https://schema.org/SequenceRange	254
pairedomicsdata.bioinformatics.nl	https://schema.org/Person	222
www.ebi.ac.uk	http://xmlns.com/foaf/0.1/Image	222
proteinensemble.org	https://schema.org/CollectionPage	187
proteinensemble.org	https://schema.org/CreativeWork	187
proteinensemble.org	https://schema.org/DefinedTermSet	187
proteinensemble.org	https://schema.org/ItemList	187
proteinensemble.org	https://schema.org/ScholarlyArticle	181
pairedomicsdata.bioinformatics.nl	https://schema.org/ContactPoint	148
www.covid19dataportal.org	https://schema.org/Organization	148
www.covid19dataportal.org	https://schema.org/Dataset	110
www.france-bioinformatique.fr	https://schema.org/BreadcrumbList	99
	https://schema.org/Dataset	97
test.intermine.org	https://schema.org/Protein	97
www.ebi.ac.uk	http://xmlns.com/foaf/0.1/Document	91
	https://schema.org/Person	90
bgee.org	https://schema.org/Dataset	87
pairedomicsdata.bioinformatics.nl	https://schema.org/CreativeWork	74
pairedomicsdata.bioinformatics.nl	https://schema.org/DataCatalog	74
pairedomicsdata.bioinformatics.nl	https://schema.org/DataDownload	74
pairedomicsdata.bioinformatics.nl	https://schema.org/Dataset	74
pairedomicsdata.bioinformatics.nl	https://schema.org/Organization	74
www.ebi.ac.uk	http://rdfs.org/sioc/ns#Item	59
www.france-bioinformatique.fr	https://schema.org/Event	40
www.france-bioinformatique.fr	https://schema.org/Place	40
www.france-bioinformatique.fr	https://schema.org/contact	40
www.france-bioinformatique.fr	https://schema.org/hostInstitution	40
test.intermine.org	https://schema.org/Gene	39
	https://bioschemas.org/Taxon	29
nanocommons.github.io	https://schema.org/ChemicalSubstance	29
	https://schema.org/Organization	27
bridgedb.github.io	https://schema.org/DataDownload	23
bridgedb.github.io	https://schema.org/Dataset	23
www.covid19dataportal.org	https://schema.org/DataCatalog	19
www.uniprot.org	https://schema.org/Organization	14
	https://schema.org/CreativeWork	9
nanocommons.github.io	https://schema.org/Organization	9
bgee.org	https://schema.org/Answer	8
bgee.org	https://schema.org/CreativeWork	8
bgee.org	https://schema.org/Question	8

domain	type	count
	https://schema.org/WebPage	7
nanocommons.github.io	https://schema.org/Dataset	7
nanocommons.github.io	https://schema.org/DataDownload	6
	https://schema.org/DataCatalog	5
bgee.org	https://schema.org/Offer	5
	https://schema.org/SearchAction	4
bgee.org	https://schema.org/SoftwareSourceCode	4
nanocommons.github.io	https://schema.org/CreativeWork	4
	https://schema.org/EducationalOrganization	3
	https://schema.org/NGO	3
	https://schema.org/ScholarlyArticle	3
bgee.org	https://schema.org/WebApplication	3
biopragmatics.github.io	https://schema.org/Person	3
prosite.expasy.org	https://schema.org/Action	3
	https://schema.org/CollegeOrUniversity	2
	https://schema.org/WebSite	2
bgee.org	https://schema.org/SoftwareApplication	2
biopragmatics.github.io	https://schema.org/WebPage	2
bridgedb.github.io	https://schema.org/CreativeWork	2
www.uniprot.org	https://schema.org/GovernmentOrganization	2
www.uniprot.org	https://schema.org/NGO	2
www.uniprot.org	https://schema.org/WebPage	2
	https://schema.org/GovernmentOrganization	1
bgee.org	https://schema.org/AboutPage	1
bgee.org	https://schema.org/FAQPage	1
biopragmatics.github.io	https://schema.org/WebSite	1
bridgedb.github.io	https://schema.org/SoftwareApplication	1
bridgedb.github.io	https://schema.org/WebPage	1
bridgedb.github.io	https://schema.org/WebSite	1
massbank.eu	https://schema.org/DataCatalog	1
massbank.eu	https://schema.org/ScholarlyArticle	1
nanocommons.github.io	https://schema.org/DataCatalog	1
nanocommons.github.io	https://schema.org/Person	1
nanocommons.github.io	https://schema.org/URL	1
nanocommons.github.io	https://schema.org/WebSite	1
prosite.expasy.org	https://schema.org/DigitalDocument	1
prosite.expasy.org	https://schema.org/SearchAction	1
www.metanetx.org	https://schema.org/SoftwareApplication	1
www.uniprot.org	https://schema.org/Dataset	1
www.uniprot.org	https://schema.org/QAPage	1

Connectivity of the Data

We were interested to gain some insight as to how connected the data was both internally, and how many points where it would link up with other knowledge graphs. The queries in this section focus on the connectedness of the data.

We first investigated the number of nodes that only contained incoming edges. We report the total number of object nodes there are (excluding literals), and the number of edge IRIs, i.e. those that only have incoming properties. Only 4.65% of the nodes only contain incoming edges.

Object IRIs	Edge IRIs
2,057,094	95,610

We then investigated the number of outgoing links per class. We report here the top 20 results.

s	class	nb_out_edges
https://bioschemas.org/crawl/v1/bgee/?page=gene&gene_id=ENSBTAG00000027937/20211110/90020/bgee.org/?page=gene&gene_id=ENSBTAG00000027937/1779564251	https://bioschemas.org/Gene	856
https://www.metanetx.org/chem_info/MNXM1944	https://schema.org/MolecularEntity	654
https://doi.org/10.15468/hb9rjv	https://schema.org/Dataset	594
https://bioschemas.org/crawl/v1/bgee/?page=gene&gene_id=ENSBTAG00000043564/20211110/94715/bgee.org/?page=gene&gene_id=ENSBTAG00000043564/1772156424	https://bioschemas.org/Gene	519
https://bioschemas.org/crawl/v1/bgee/?page=gene&gene_id=ENSBTAG00000043584/20211110/94734/bgee.org/?page=gene&gene_id=ENSBTAG00000043584/2022662406	https://bioschemas.org/Gene	474
https://doi.org/10.15468/m5vrza	https://schema.org/Dataset	406
http://www.ebi.ac.uk/pdbe/about/past-events	http://rdfs.org/sioc/ns#Item	346
http://www.ebi.ac.uk/pdbe/about/past-events	http://xmlns.com/foaf/0.1/Document	346
https://doi.org/10.15468/vmf5ye	https://schema.org/Dataset	296
https://bioschemas.org/crawl/v1/bgee/?page=gene&gene_id=ENSBTAG00000043559/20211110/94710/bgee.org/?page=gene&gene_id=ENSBTAG00000043559/1377128066	https://bioschemas.org/Gene	292
https://doi.org/10.5281/zenodo.291971	https://schema.org/Dataset	289
https://bioschemas.org/crawl/v1/bgee/?page=gene&gene_id=ENSBTAG00000043546/20211110/94697/bgee.org/?page=gene&gene_id=ENSBTAG00000043546/1804549344	https://bioschemas.org/Gene	284
https://doi.org/10.15472/hy9nif	https://schema.org/Dataset	282
https://bioschemas.org/crawl/v1/bgee/?page=gene&gene_id=ENSBTAG00000043550/20211110/94701/bgee.org/?page=gene&gene_id=ENSBTAG00000043550/1476242225	https://bioschemas.org/Gene	269
https://www.metanetx.org/chem_info/MNXM383	https://schema.org/MolecularEntity	264

BioHackrXiv

https://bioschemas.org/crawl/v1/bgee/?page=gene&gene_id=ENSBTAG00000043577/20211110/94727/bgee.org/?page=gene&gene_id=ENSBTAG00000043577/1610681495	https://bioschemas.org/Gene	261
https://bioschemas.org/crawl/v1/bgee/?page=gene&gene_id=ENSBTAG00000043556/20211110/94707/bgee.org/?page=gene&gene_id=ENSBTAG00000043556/1277162978	https://bioschemas.org/Gene	240
https://bioschemas.org/crawl/v1/bgee/?page=gene&gene_id=ENSBTAG00000043568/20211110/94719/bgee.org/?page=gene&gene_id=ENSBTAG00000043568/2065005542	https://bioschemas.org/Gene	235
https://bioschemas.org/crawl/v1/bgee/?page=gene&gene_id=ENSBTAG00000043560/20211110/94711/bgee.org/?page=gene&gene_id=ENSBTAG00000043560/1818154049	https://bioschemas.org/Gene	229

Discussion

Through this Hackathon project we have demonstrated that it is possible to harvest Bioschemas markup from a number of sites and load them into a triplestore. However, this has revealed a number of challenges that need to be resolved.

Harvesting content from a whole site is very time consuming, particularly for dynamic sites. Harvesting requires visiting each page with markup in turn and extracting the markup. In the case of dynamic sites the content needs to be rendered before it can be extracted. Most of the sites that could be completed harvested did not contain data content beyond Dataset and DataCatalog.

The quality of deployed markup is very problematic. As reported above, a number of sites could not be harvested due to issues with their sitemap. Of those that could be harvested, a number of pages were not harvested due to issues with the markup contained within them, e.g. inclusion of non-escaped characters within strings was a common error. Even the markup that could be extracted contained errors, e.g. the use of different namespaces for the declaration of the Gene type as identified in the Instances per Bioschemas Class query. This highlights the need for a Bioschemas validator capable of both syntactic and semantic checking (see the data validation outputs of Project 29).

Future work

The next steps for this work would be to improve the robustness of the data harvesting pipeline, including automating the manual steps of iterating over an index of sitemap files; merging individual files for each harvested page into a single file per subsitemap; and loading the harvested data into the triplestore.

The use of data dumps for sites should be considered to eliminate the need for data harvesting, which is a costly process for both data producers – who have to have sufficient bandwidth and compute resources to serve the content – and data consumers – who have to have sufficient compute resources to retrieve, render, and extract the content from each page.

Jupyter notebooks, GitHub repositories and data repositories

- GitHub repository: https://github.com/BioSchemas/bioschemas-data-harvesting
- Jupyter Notebook: https://github.com/BioSchemas/bioschemas-data-harvesting/blob/main/AnalysisQueries.ipynb
  - MyBinder launch: https://mybinder.org/v2/gh/BioSchemas/bioschemas-data-harvesting/HEAD?labpath=AnalysisQueries.ipynb
- SPARQL Endpoint: https://swel.macs.hw.ac.uk/data/repositories/bioschemas
  - Snorql Extended Interface: https://swel.macs.hw.ac.uk/bioschemas/
- Data download director: https://swel.macs.hw.ac.uk/bioschemas-data/

Acknowledgements

This work was done during the BioHackathon Europe 2021 organised by ELIXIR and run in November 2021. We wish to thank the organizers and supporters of the Biohackathon Europe 2021 for offering the venue for improving Bioschemas community efforts.

References

Dumontier, M., Gray, A. J., Marshall, M. S., Alexiev, V., Ansell, P., Bader, G., Baran, J., Bolleman, J. T., Callahan, A., Cruz-Toledo, J.others. (2016). The health care and life sciences community profile for dataset descriptions. *PeerJ*, *4*, e2331. https://doi.org/10.7717/peerj.2331

Gray, A. J. G., Papadopoulos, P., Asif, I., Mičetić, I., & Hatos, A. (2022, January 11). Creating and exploiting the intrinsically disordered protein knowledge graph (IDP-KG). *13th International Semantic Web Applications and Tools for Health Care and Life Sciences Conference (Swat4hcls 2022)*. http://www.swat4ls.org/

Gray, A. J. G., Papadopoulos, P., Mičetić, I., & Hatos, A. (2021). *Exploiting bioschemas markup to populate IDPcentral*. BioHackrXiv. https://doi.org/10.37044/osf.io/v3jct