Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche

ISTI Technical Reports

Bioschemas data sources aggregation to OpenAIRE Research Graph

Enrico Ottonello, ISTI-CNR, Pisa, Italy
Michele Artini, ISTI-CNR, Pisa, Italy
Sandro La Bruzzo, ISTI-CNR, Pisa, Italy
Gina Pavone, ISTI-CNR, Pisa, Italy

ISTI-TR-2022/010

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche

Bioschemas data sources aggregation to OpenAIRE Research Graph
Ottonello E., Artini M., La Bruzzo S., Pavone G.
ISTI-TR-2022/010

In this report we propose an extended Hadoop-based aggregator for the harvesting of Bioschemas data sources. In this extended hadoop-based aggregator, the downloaded data will be processed according to the consolidated data flow: the original contents will be mapped onto an internal representation that will make them eligible to be integrated in the OpenAIRE research graph.

Keywords: Bioschemas, OpenAIRE, Data science, Data integration, Open science.

Citation
Ottonello E., Artini M., La Bruzzo S., Pavone G., Bioschemas data sources aggregation to OpenAIRE Research Graph. ISTI Technical Reports 2022/010. DOI: 10.32079/ISTI-TR-2022/010.

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Area della Ricerca CNR di Pisa
Via G. Moruzzi 1
56124 Pisa Italy
http://www.isti.cnr.it

Bioschemas data sources aggregation to OpenAIRE Research Graph

Version: 1.0 **Release Date**: 26th of April, 2022

Authors: Enrico Ottonello, Michele Artini, Sandro La Bruzzo, Gina Pavone

1 Institute of information science and technologies - CNR, Pisa, Italy

Introduction

Bioschemas (https://bioschemas.org/index.html) is a community project built on top of schema.org, aiming to improve interoperability and findability on the Web of Life Sciences resources such as datasets, software, and training materials. It allows resources to better communicate and work together by using a common markup on their websites.
Moreover, it encourages people in the life sciences to use Schema.org markup in their websites so that they are indexable by search engines and other services.
The goal of this report is to show how to extend the capabilities of the metadata aggregation subsystem, inside the context of OpenAIRE information graph (https://explore.openaire.eu/), in order to harvest Bioschemas data sources.
The already existing procedures in the OpenAIRE Hadoop-based aggregator include plugins for OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) or leveraging REST APIs. However, the kind of (meta)data providers we have to address in Bioschemas context are websites implemented according to some guidelines. Those websites expose the bibliographic information in a way that can be collected using the tool BMUSE: Bioschemas Markup Scraper and Extractor (https://github.com/HW-SWeL/BMUSE). This tool was selected in collaboration with the EOSC-Enhance partners for the harvesting of such information.
In this report we propose an extended Hadoop-based aggregator for the harvesting of Bioschemas data sources. In this extended hadoop-based aggregator, the downloaded data will be processed according to the consolidated data flow: the original contents will be mapped onto an internal representation that will make them eligible to be integrated in the OpenAIRE research graph.
In the following of this report we show before the considered data sources, then the architecture of the extended hadoop-based aggregator, and its implemented workflows. In the description of the workflows we refer to a specific record of a Bioschemas Protein profile.

Bioschemas data sources

The complete list of existing bioschemas data sources is available on Bioschemas website (https://bioschemas.org/liveDeploys). To simplify the marking up of web resources, and to provide consistency of markup within the life sciences community (https://bioschemas.org/index.html), Bioschemas provides 'profiles', which are customisation of schema.org types including important guidelines on how to use it within the Life Sciences domain. A profile can be used to define the semantics of a particular property, the valid value(s) and ranges that may be attributed to that property, and the cardinality with which that property may appear.
In this report, we refer to a specific Bioschemas data source related to Bioschemas Protein profile (https://bioschemas.org/profiles/Protein/0.11-RELEASE), and we consider the most interesting harvested data source from Protein Ensemble (https://proteinensemble.org/) because its markup also contains information related to the citations.
The image below (Fig 1) shows an example of website (https://proteinensemble.org/PED00001) implemented according to Bioschemas guidelines

(https://bioschemas.org/tutorials/howto/howto_add_markup) describing a structural ensemble, including intrinsically disordered proteins (IDPs). The HTML source contains the JSON-LD Bioschemas markup.

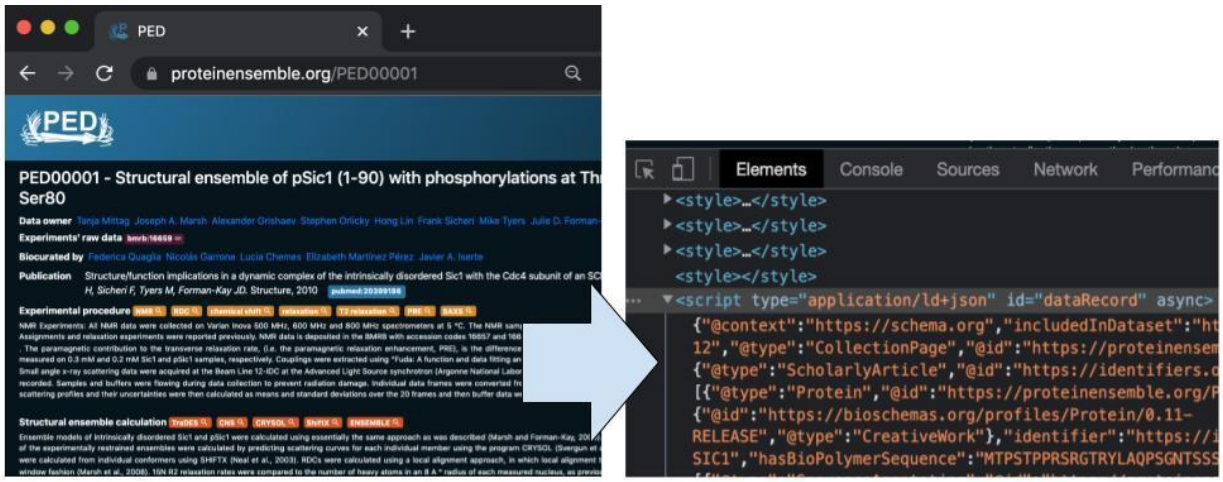

Figure 1 Information extraction from Bioschemas web site

Below there is the complete JSON-LD Bioschemas markup found in the HTML source of the website:


```
{
   "@context": "https://schema.org",
   "includedInDataset": "https://proteinensemble.org/#2021-02-12",
   "@type": "CollectionPage",
   "@id": "https://proteinensemble.org/PED00001",
   "citation": {
      "@type": "ScholarlyArticle",
      "@id": "https://identifiers.org/pubmed:20399186"
   },
   "mainEntity": {
      "@type": "ItemList",
      "numberOfItems": 1,
      "itemListElement": [
         {
            "@type": "Protein",
            "@id": "https://proteinensemble.org/PED00001#P38634_A_1",
            "http://purl.org/dc/terms/conformsTo": {
               "@id": "https://bioschemas.org/profiles/Protein/0.11-RELEASE",
               "@type": "CreativeWork"
            },
            "identifier": "https://identifiers.org/uniprot:P38634",
            "sameAs": "http://purl.uniprot.org/uniprot/P38634",
            "name": "Protein SIC1",
            "hasBioPolymerSequence":
"MTPSTPPRSRGTRYLAQPSGNTSSSALMQGQKTPQKPSQNLVPVTPSTTKSFKNAPLLAPPNSNMGMTSPFNGLTSPQRSPFPKSSVKRT",
            "hasSequenceAnnotation": [
               {
                  "@type": "SequenceAnnotation",
                  "@id": "https://proteinensemble.org/PED00001#P38634_A_1_1_90",
                  "sequenceLocation": {
                     "@type": "SequenceRange",
                     "rangeStart": 1,
                     "rangeEnd": 90
                  },
                  "additionalProperty": [
                     {
                        "@type": "PropertyValue",
                        "name": "Term",
                        "value": {
                           "@type": "DefinedTerm",
                           "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00120",
                           "inDefinedTermSet": {
                              "@type": "DefinedTermSet",
                              "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl",
                              "name": "IDP ontology"
                           },
                           "termCode": "IDPO:00120",
```

```
            "name": "NMR"
         }
      },
      {
         "@type": "PropertyValue",
         "name": "Term",
         "value": {
            "@type": "DefinedTerm",
            "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00166",
            "inDefinedTermSet": {
               "@type": "DefinedTermSet",
               "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl",
               "name": "IDP ontology"
            },
            "termCode": "IDPO:00166",
            "name": "RDC"
         }
      },
      {
         "@type": "PropertyValue",
         "name": "Term",
         "value": {
            "@type": "DefinedTerm",
            "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00167",
            "inDefinedTermSet": {
               "@type": "DefinedTermSet",
               "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl",
               "name": "IDP ontology"
            },
            "termCode": "IDPO:00167",
            "name": "chemical shift"
         }
      },
      {
         "@type": "PropertyValue",
         "name": "Term",
         "value": {
            "@type": "DefinedTerm",
            "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00168",
            "inDefinedTermSet": {
               "@type": "DefinedTermSet",
               "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl",
               "name": "IDP ontology"
            },
            "termCode": "IDPO:00168",
            "name": "relaxation"
         }
      },
      {
         "@type": "PropertyValue",
         "name": "Term",
         "value": {
            "@type": "DefinedTerm",
            "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00169",
            "inDefinedTermSet": {
               "@type": "DefinedTermSet",
               "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl",
               "name": "IDP ontology"
            },
            "termCode": "IDPO:00169",
            "name": "T2 relaxation"
         }
      },
      {
         "@type": "PropertyValue",
         "name": "Term",
         "value": {
            "@type": "DefinedTerm",
            "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00171",
            "inDefinedTermSet": {
               "@type": "DefinedTermSet",
               "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl",
               "name": "IDP ontology"
            },
            "termCode": "IDPO:00171",
            "name": "PRE "
         }
      },
      {
         "@type": "PropertyValue",
         "name": "Term",
         "value": {
            "@type": "DefinedTerm",
            "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00125",
            "inDefinedTermSet": {
               "@type": "DefinedTermSet",
               "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl",
               "name": "IDP ontology"
            },
            "termCode": "IDPO:00125",
            "name": "SAXS"
         }
      },
```

```
                        {
                           "@type": "PropertyValue",
                           "name": "Term",
                           "value": {
                              "@type": "DefinedTerm",
                              "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl:00186",
                              "inDefinedTermSet": {
                                 "@type": "DefinedTermSet",
                                 "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl",
                                 "name": "IDP ontology"
                              },
                              "termCode": "IDPO:00186",
                              "name": "TraDES"
                           }
                        },
                        {
                           "@type": "PropertyValue",
                           "name": "Term",
                           "value": {
                              "@type": "DefinedTerm",
                              "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl:00192",
                              "inDefinedTermSet": {
                                 "@type": "DefinedTermSet",
                                 "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl",
                                 "name": "IDP ontology"
                              },
                              "termCode": "IDPO:00192",
                              "name": "CNS"
                           }
                        },
                        {
                           "@type": "PropertyValue",
                           "name": "Term",
                           "value": {
                              "@type": "DefinedTerm",
                              "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl:00208",
                              "inDefinedTermSet": {
                                 "@type": "DefinedTermSet",
                                 "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl",
                                 "name": "IDP ontology"
                              },
                              "termCode": "IDPO:00208",
                              "name": "CRYSOL"
                           }
                        },
                        {
                           "@type": "PropertyValue",
                           "name": "Term",
                           "value": {
                              "@type": "DefinedTerm",
                              "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl:00210",
                              "inDefinedTermSet": {
                                 "@type": "DefinedTermSet",
                                 "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl",
                                 "name": "IDP ontology"
                              },
                              "termCode": "IDPO:00210",
                              "name": "ShiftX"
                           }
                        },
                        {
                           "@type": "PropertyValue",
                           "name": "Term",
                           "value": {
                              "@type": "DefinedTerm",
                              "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl:00216",
                              "inDefinedTermSet": {
                                 "@type": "DefinedTermSet",
                                 "@id": "https://disprot.org/assets/data/IDPO_v0.2.owl",
                                 "name": "IDP ontology"
                              },
                              "termCode": "IDPO:00216",
                              "name": "ENSEMBLE"
                           }
                        }
                     ]
                  }
               ]
            }
         ]
      },
      "identifier": "https://identifiers.org/ped:PED00001",
      "name": "Structural ensemble of pSic1 (1-90) with phosphorylations at Thr5, Thr33, Thr45, Ser69, Ser76, Ser80"
}
```

Extended hadoop-based aggregator

In this section, we describe:

- the reference architecture of the extended hadoop-based aggregator;
- the aggregation workflows.

Architecture

The extended hadoop-based aggregator exploits existing technologies such as Oozie, Spark, Hadoop and BMUSE. In the following, a brief description of these technologies is provided.
Figure 2 shows a reference architecture of the extended hadoop-based aggregator.
The code has been developed using Java language.

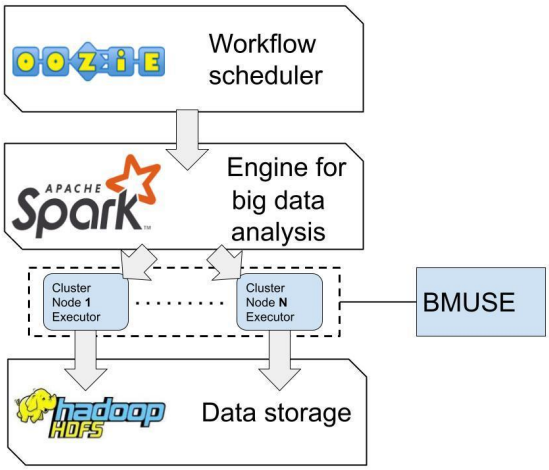

Figure 2 Extended hadoop-based aggregator

Spark

Apache Spark is a unified analytics engine for large-scale data processing (https://spark.apache.org/). It provides high-level APIs in Java, Scala, Python and R, and an optimized engine that supports general execution graphs. It also supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing, MLlib for machine learning, GraphX for graph processing, and Structured Streaming for incremental computation and stream processing.

Hadoop

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing (https://hadoop.apache.org/).
The Apache Hadoop software library is a framework that allows for the distributed processing of large datasets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to

detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

Oozie

Oozie is a workflow scheduler system to manage Apache Hadoop jobs (https://oozie.apache.org/).
Oozie Coordinator jobs are recurrent Oozie Workflow jobs triggered by time (frequency) and data availability.
Oozie is integrated with the rest of the Hadoop stack supporting several types of Hadoop jobs out of the box (such as Java map-reduce, Streaming map-reduce, Pig, Hive, Sqoop and Distcp) as well as system specific jobs (such as Java programs and shell scripts).
Oozie is a scalable, reliable and extensible system.

BMUSE tool

BMUSE (https://github.com/HW-SWeL/BMUSE) is a directed scraping framework capable of extracting JSON-LD and RDFa markup from static and single page application sites, in our case the markup is returned in N-Quads format, a line-based, plain text format for encoding an RDF dataset (https://www.w3.org/TR/n-quads/).RDF is a standard model for data interchange on the Web (https://www.w3.org/RDF/).

Aggregation workflows

Each of the following steps was implemented as an oozie workflow, to avoid conflicts among those main components:

- BMUSE tool
- RDF libraries (rdf4j)
- OpenAIRE framework on hadoop

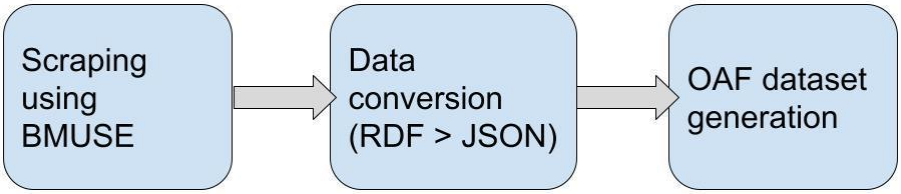

Figure 3 Workflows sequence

Scraping workflow

Scraping workflow defines a java action that run on a single worker node and it is responsible to collect the bibliographic records from a single website (modeled in OpenAIRE as a data source) sequentially, storing them on HDFS.

Input

The input is the sitemap, that can be compressed or not, of a single data source website, like this one: https://proteinensemble.org/sitemap2.xml.gz, containing all the urls of the webpages each one describing a Bioschemas resource.

Output

The output is a sequence file on HDFS containing N-Quads format entries representing the scraped markups.

N-Quads example related to a single scraped page:

```
https://proteinensemble.org/PED00001  <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> <http://purl.org/pav/retrievedFrom>
<https://proteinensemble.org/PED00001> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> <http://purl.org/pav/retrievedOn>
"2021-12-10T11:11:09"^^<http://www.w3.org/2001/XMLSchema#dateTime> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> <http://purl.org/pav/createdWith> <https://github.com/HW-SWeL/BMUSE/releases/tag/0.5.2> .
<https://proteinensemble.org/PED00001> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/CollectionPage>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001> <https://schema.org/citation> <https://identifiers.org/pubmed:20399186>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001> <https://schema.org/identifier> "https://identifiers.org/ped:PED00001"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001> <https://schema.org/includedInDataset> "https://proteinensemble.org/#2021-02-12"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001> <https://schema.org/mainEntity>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1965306778>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001> <https://schema.org/name> "Structural ensemble of pSic1 (1-90) with phosphorylations at Thr5, Thr33, Thr45, Ser69, Ser76,
Ser80" <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://identifiers.org/pubmed:20399186> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/ScholarlyArticle>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1965306778>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/ItemList> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1965306778> <https://schema.org/itemListElement>
<https://proteinensemble.org/PED00001#P38634_A_1> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1965306778> <https://schema.org/numberOfItems> "1"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/Protein>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1> <http://purl.org/dc/terms/conformsTo> <https://bioschemas.org/profiles/Protein/0.11-RELEASE>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1> <https://schema.org/hasBioPolymerSequence>
"MTPSTPPRSRGTRYLAQPSGNTSSSALMQGQKTPQKPSQNLVPVTPSTTKSFKNAPLLAPPNSNMGMTSPFNGLTSPQRSPFPKSSVKRT"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1> <https://schema.org/hasSequenceAnnotation> <https://proteinensemble.org/PED00001#P38634_A_1_1_90>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1> <https://schema.org/identifier> "https://identifiers.org/uniprot:P38634"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1> <https://schema.org/name> "Protein SIC1"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1> <https://schema.org/sameAs> <http://purl.uniprot.org/uniprot/P38634>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/profiles/Protein/0.11-RELEASE> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/CreativeWork>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1_1_90> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/SequenceAnnotation>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1_1_90> <https://schema.org/additionalProperty>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/443479062>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1_1_90> <https://schema.org/additionalProperty>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/834145387>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1_1_90> <https://schema.org/additionalProperty>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1180239248>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1_1_90> <https://schema.org/additionalProperty>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/196727103>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
```

<https://proteinensemble.org/PED00001#P38634_A_1_1_90> <https://schema.org/additionalProperty>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1585199578>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1_1_90> <https://schema.org/additionalProperty>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1178902959>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1_1_90> <https://schema.org/additionalProperty>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1532438514>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1_1_90> <https://schema.org/additionalProperty>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/663212760>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1_1_90> <https://schema.org/additionalProperty>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/225227635>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1_1_90> <https://schema.org/additionalProperty>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/404264600>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1_1_90> <https://schema.org/additionalProperty>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/872740108>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1_1_90> <https://schema.org/additionalProperty>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/831429981>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001#P38634_A_1_1_90> <https://schema.org/sequenceLocation>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/414547497>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/443479062>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/PropertyValue> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/443479062> <https://schema.org/name> "Term"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/443479062> <https://schema.org/value>
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00120> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00120> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/DefinedTerm>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00120> <https://schema.org/inDefinedTermSet> <https://disprot.org/assets/data/IDPO_v0.2.owl>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00120> <https://schema.org/name> "NMR"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00120> <https://schema.org/termCode> "IDPO:00120"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/DefinedTermSet>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl> <https://schema.org/name> "IDP ontology" <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/834145387>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/PropertyValue> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/834145387> <https://schema.org/name> "Term"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/834145387> <https://schema.org/value>
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00166> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00166> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/DefinedTerm>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00166> <https://schema.org/inDefinedTermSet> <https://disprot.org/assets/data/IDPO_v0.2.owl>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00166> <https://schema.org/name> "RDC"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00166> <https://schema.org/termCode> "IDPO:00166"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1180239248>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/PropertyValue> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1180239248> <https://schema.org/name> "Term"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1180239248> <https://schema.org/value>
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00167> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00167> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/DefinedTerm>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00167> <https://schema.org/inDefinedTermSet> <https://disprot.org/assets/data/IDPO_v0.2.owl>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00167> <https://schema.org/name> "chemical shift"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00167> <https://schema.org/termCode> "IDPO:00167"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/196727103>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/PropertyValue> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/196727103> <https://schema.org/name> "Term"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/196727103> <https://schema.org/value>
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00168> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00168> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/DefinedTerm>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .

<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00168> <https://schema.org/inDefinedTermSet> <https://disprot.org/assets/data/IDPO_v0.2.owl>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00168> <https://schema.org/name> "relaxation"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00168> <https://schema.org/termCode> "IDPO:00168"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1585199578>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/PropertyValue> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1585199578> <https://schema.org/name> "Term"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1585199578> <https://schema.org/value>
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00169> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00169> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/DefinedTerm>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00169> <https://schema.org/inDefinedTermSet> <https://disprot.org/assets/data/IDPO_v0.2.owl>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00169> <https://schema.org/name> "T2 relaxation"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00169> <https://schema.org/termCode> "IDPO:00169"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1178902959>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/PropertyValue> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1178902959> <https://schema.org/name> "Term"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1178902959> <https://schema.org/value>
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00171> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00171> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/DefinedTerm>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00171> <https://schema.org/inDefinedTermSet> <https://disprot.org/assets/data/IDPO_v0.2.owl>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00171> <https://schema.org/name> "PRE "
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00171> <https://schema.org/termCode> "IDPO:00171"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1532438514>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/PropertyValue> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1532438514> <https://schema.org/name> "Term"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/1532438514> <https://schema.org/value>
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00125> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00125> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/DefinedTerm>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00125> <https://schema.org/inDefinedTermSet> <https://disprot.org/assets/data/IDPO_v0.2.owl>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00125> <https://schema.org/name> "SAXS"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00125> <https://schema.org/termCode> "IDPO:00125"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/663212760>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/PropertyValue> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/663212760> <https://schema.org/name> "Term"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/663212760> <https://schema.org/value>
<https://disprot.org/assets/data/IDPO_v0.2.owl:00186> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00186> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/DefinedTerm>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00186> <https://schema.org/inDefinedTermSet> <https://disprot.org/assets/data/IDPO_v0.2.owl>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00186> <https://schema.org/name> "TraDES" <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00186> <https://schema.org/termCode> "IDPO:00186"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/225227635>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/PropertyValue> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/225227635> <https://schema.org/name> "Term"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/225227635> <https://schema.org/value>
<https://disprot.org/assets/data/IDPO_v0.2.owl:00192> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00192> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/DefinedTerm>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00192> <https://schema.org/inDefinedTermSet> <https://disprot.org/assets/data/IDPO_v0.2.owl>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00192> <https://schema.org/name> "CNS" <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00192> <https://schema.org/termCode> "IDPO:00192"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/404264600>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/PropertyValue> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/404264600> <https://schema.org/name> "Term"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/404264600> <https://schema.org/value>
<https://disprot.org/assets/data/IDPO_v0.2.owl:00208> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .

<https://disprot.org/assets/data/IDPO_v0.2.owl:00208> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/DefinedTerm>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00208> <https://schema.org/inDefinedTermSet> <https://disprot.org/assets/data/IDPO_v0.2.owl>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00208> <https://schema.org/name> "CRYSOL" <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00208> <https://schema.org/termCode> "IDPO:00208"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/872740108>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/PropertyValue> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/872740108> <https://schema.org/name> "Term"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/872740108> <https://schema.org/value>
<https://disprot.org/assets/data/IDPO_v0.2.owl:00210> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00210> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/DefinedTerm>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00210> <https://schema.org/inDefinedTermSet> <https://disprot.org/assets/data/IDPO_v0.2.owl>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00210> <https://schema.org/name> "ShiftX" <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00210> <https://schema.org/termCode> "IDPO:00210"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/831429981>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/PropertyValue> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/831429981> <https://schema.org/name> "Term"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/831429981> <https://schema.org/value>
<https://disprot.org/assets/data/IDPO_v0.2.owl:00216> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00216> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/DefinedTerm>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00216> <https://schema.org/inDefinedTermSet> <https://disprot.org/assets/data/IDPO_v0.2.owl>
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00216> <https://schema.org/name> "ENSEMBLE"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://disprot.org/assets/data/IDPO_v0.2.owl:00216> <https://schema.org/termCode> "IDPO:00216"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/414547497>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://schema.org/SequenceRange> <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/414547497> <https://schema.org/rangeEnd> "90"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0/proteinensemble.org/PED00001/414547497> <https://schema.org/rangeStart> "1"
<https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .
<https://proteinensemble.org/PED00001> <http://purl.org/dc/terms/title> "PED" <https://bioschemas.org/crawl/v1/proteinensemble/PED00001/20211210/0> .

Data conversion (RDF > JSON) workflow

The data conversion workflow applies the proper data mapping using a spark action, thus better exploiting the parallelism of cluster resources, and generates on HDFS JSON data.

rdf-converter-cmdline tool

This workflow is based on the experience that we acquired developing the rdf-converter-cmdline tool (https://github.com/openaire/rdf-converter-cmdline), a Java-based command line software with the aim of converting rdf files from N-Quads format to JSON files following datacite schema (https://github.com/datacite/schema/blob/master/source/json/kernel-4.3/datacite_4.3_schema.json). This tool takes as input a folder containing N-Quads file generated by BMUSE.

Data models

The model representing the data from bioschemas input source was created using jackson annotations, mapping each relevant field. Below are a few examples related to this model.

BioSchemaProtein model:

Property	Type	Multiplicity	Jackson JsonProperty Annotation
id	String	1	@id
entryList	Entry	1..N	@graph
retrievedOn	DateTimeType	1	http://purl.org/pav/retrievedOn

A few properties from Entry model:

Property	Type	Multiplicity	Jackson JsonProperty Annotation
id	String	1	@id
type	String	1	@type
identifier	String	1	https://schema.org/identifier
name	String	1	https://schema.org/name
url	String	1	url
alternateName	String	1	alternateName
citation	Citation	1	https://schema.org/citation
sameAs	Link	1..N	https://schema.org/sameAs

Citation model:

Property	Type	Multiplicity	Jackson JsonProperty Annotation
id	String	1	@id

The model representing the Protein was created mapping all the fields that could be retrieved from the bioschemas input source to an existing field of the Datacite schema. Below there are some examples related to this model.

A few properties from DataciteProtein model:

Property	Type	Multiplicity
id	String	1
types	Types	1..N
titles	Title	1..N

Title model:

Property	Type	Multiplicity
title	String	1
titleType	String	1

Types model:

Property	Type	Multiplicity
resourceType	String	1
resourceTypeGeneral	String	1

Input

The workflow takes as input the sequence file that contains the N-Quads data stored on HDFS.

Output

The result of the conversion is a set of JSON records according to Datacite format, like in the example below:


```
{
   "id": "PED00001#P38634_A_1",
   "types": {
      "resourceType": "Protein",
      "resourceTypeGeneral": "Dataset"
   },
   "creators": [],
   "identifiers": [
      {
         "identifier": "https://proteinensemble.org/PED00001#P38634_A_1",
         "identifierType": "URL"
      }
   ],
   "relatedIdentifiers": [
      {
         "relationType": "IsCitedBy",
         "relatedIdentifier": "https://identifiers.org/pubmed:20399186",
         "relatedIdentifierType": "URL"
      },
      {
```

```
         "relationType": "IsIdenticalTo",
         "relatedIdentifier": "http://purl.uniprot.org/uniprot/P38634",
         "relatedIdentifierType": "URL"
      }
   ],
   "alternateIdentifiers": [
      {
         "alternateIdentifier": "https://identifiers.org/uniprot:P38634"
      }
   ],
   "descriptions": [],
   "titles": [
      {
         "title": "PED00001#P38634_A_1 - Structural ensemble of pSic1 (1-90) with phosphorylations at Thr5, Thr33, Thr45,
Ser69, Ser76, Ser80"
      }
   ],
   "dates": [
      {
         "date": "2021-12-10T11:11:09",
         "dateType": "Collected"
      }
   ],
   "subjects": [
      {
         "schemeURI": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00120",
         "value": "NMR",
         "subjectScheme": "IDPO:00120"
      },
      {
         "schemeURI": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00166",
         "value": "RDC",
         "subjectScheme": "IDPO:00166"
      },
      {
         "schemeURI": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00167",
         "value": "chemical shift",
         "subjectScheme": "IDPO:00167"
      },
      {
         "schemeURI": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00168",
         "value": "relaxation",
         "subjectScheme": "IDPO:00168"
      },
      {
         "schemeURI": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00169",
         "value": "T2 relaxation",
         "subjectScheme": "IDPO:00169"
      },
      {
         "schemeURI": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00171",
         "value": "PRE ",
         "subjectScheme": "IDPO:00171"
      },
      {
         "schemeURI": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00125",
         "value": "SAXS",
         "subjectScheme": "IDPO:00125"
      },
      {
         "schemeURI": "https://disprot.org/assets/data/IDPO_v0.2.owl:00186",
         "value": "TraDES",
         "subjectScheme": "IDPO:00186"
      },
      {
         "schemeURI": "https://disprot.org/assets/data/IDPO_v0.2.owl:00192",
         "value": "CNS",
         "subjectScheme": "IDPO:00192"
      },
      {
         "schemeURI": "https://disprot.org/assets/data/IDPO_v0.2.owl:00208",
         "value": "CRYSOL",
         "subjectScheme": "IDPO:00208"
      },
      {
         "schemeURI": "https://disprot.org/assets/data/IDPO_v0.2.owl:00210",
         "value": "ShiftX",
         "subjectScheme": "IDPO:00210"
      },
      {
         "schemeURI": "https://disprot.org/assets/data/IDPO_v0.2.owl:00216",
         "value": "ENSEMBLE",
         "subjectScheme": "IDPO:00216"
      }
   ]
}
```

OAF dataset generation workflow

The following workflow generates the dataset that is eligible for being integrated in the OpenAIRE research graph.

Input

A set of json record stored on HDFS according to Datacite format, containing the informations related to the Bioschemas Protein profile harvested from a single datasource.

Output

A dataset stored on HDFS that contains entities according to OpenAIRE format; those entities are of two different types: result and relation among results.

Result example:


```
{
  "size": null,
  "geolocation": null,
  "dataInfo": {
    "provenanceaction": {
      "classid": "sysimport:actionset",
      "classname": "sysimport:actionset",
      "schemeid": "dnet:provenanceActions",
      "schemename": "dnet:provenanceActions"
    },
    "deletedbyinference": false,
    "inferred": false,
    "inferenceprovenance": null,
    "invisible": false,
    "trust": "0.9"
  },
  "resourcetype": {
    "classid": "protein",
    "classname": "protein",
    "schemeid": "dnet:publication_resource",
    "schemename": "dnet:publication_resource"
  },
  "pid": [
    {
      "dataInfo": {
        "provenanceaction": {
          "classid": "sysimport:actionset",
          "classname": "sysimport:actionset",
          "schemeid": "dnet:provenanceActions",
          "schemename": "dnet:provenanceActions"
        },
        "deletedbyinference": false,
        "inferred": false,
        "inferenceprovenance": null,
        "invisible": false,
        "trust": "0.9"
      },
      "qualifier": {
        "classid": "ped",
        "classname": "ped",
        "schemeid": "dnet:pid_types",
        "schemename": "dnet:pid_types"
      },
      "value": "PED00001#P38634_A_1"
    }
  ],
  "contributor": null,
  "oaiprovenance": null,
  "relevantdate": [
    {
      "dataInfo": null,
      "qualifier": {
        "classid": "collected",
        "classname": "collected",
        "schemeid": "dnet:dataCite_date",
        "schemename": "dnet:dataCite_date"
      },
      "value": "2021-12-10"
    }
  ],
  "collectedfrom": [
```

```
    {
      "dataInfo": {
        "provenanceaction": {
          "classid": "sysimport:actionset",
          "classname": "sysimport:actionset",
          "schemeid": "dnet:provenanceActions",
          "schemename": "dnet:provenanceActions"
        },
        "deletedbyinference": false,
        "inferred": false,
        "inferenceprovenance": null,
        "invisible": false,
        "trust": "0.9"
      },
      "key": "10|ped_________::pedDatasourceId",
      "value": "Protein Ensemble Database"
    }
  ],
  "id": "50|ped_________::c7e0527cf8940fca5c78d13c960a6d5c",
  "subject": [
    {
      "dataInfo": null,
      "qualifier": {
        "classid": "IDPO:00120",
        "classname": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00120",
        "schemeid": "dnet:subject_classification_typologies",
        "schemename": "dnet:subject_classification_typologies"
      },
      "value": "NMR"
    },
    {
      "dataInfo": null,
      "qualifier": {
        "classid": "IDPO:00166",
        "classname": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00166",
        "schemeid": "dnet:subject_classification_typologies",
        "schemename": "dnet:subject_classification_typologies"
      },
      "value": "RDC"
    },
    {
      "dataInfo": null,
      "qualifier": {
        "classid": "IDPO:00167",
        "classname": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00167",
        "schemeid": "dnet:subject_classification_typologies",
        "schemename": "dnet:subject_classification_typologies"
      },
      "value": "chemical shift"
    },
    {
      "dataInfo": null,
      "qualifier": {
        "classid": "IDPO:00168",
        "classname": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00168",
        "schemeid": "dnet:subject_classification_typologies",
        "schemename": "dnet:subject_classification_typologies"
      },
      "value": "relaxation"
    },
    {
      "dataInfo": null,
      "qualifier": {
        "classid": "IDPO:00169",
        "classname": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00169",
        "schemeid": "dnet:subject_classification_typologies",
        "schemename": "dnet:subject_classification_typologies"
      },
      "value": "T2 relaxation"
    },
    {
      "dataInfo": null,
      "qualifier": {
        "classid": "IDPO:00171",
        "classname": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00171",
        "schemeid": "dnet:subject_classification_typologies",
        "schemename": "dnet:subject_classification_typologies"
      },
      "value": "PRE "
    },
    {
      "dataInfo": null,
      "qualifier": {
        "classid": "IDPO:00125",
        "classname": "https://disprot.org/assets/data/IDPO_v0.2.owl#IDPO:00125",
        "schemeid": "dnet:subject_classification_typologies",
        "schemename": "dnet:subject_classification_typologies"
      },
      "value": "SAXS"
    },
    {
      "dataInfo": null,
      "qualifier": {
```

```
        "classid": "IDPO:00186",
        "classname": "https://disprot.org/assets/data/IDPO_v0.2.owl:00186",
        "schemeid": "dnet:subject_classification_typologies",
        "schemename": "dnet:subject_classification_typologies"
      },
      "value": "TraDES"
    },
    {
      "dataInfo": null,
      "qualifier": {
        "classid": "IDPO:00192",
        "classname": "https://disprot.org/assets/data/IDPO_v0.2.owl:00192",
        "schemeid": "dnet:subject_classification_typologies",
        "schemename": "dnet:subject_classification_typologies"
      },
      "value": "CNS"
    },
    {
      "dataInfo": null,
      "qualifier": {
        "classid": "IDPO:00208",
        "classname": "https://disprot.org/assets/data/IDPO_v0.2.owl:00208",
        "schemeid": "dnet:subject_classification_typologies",
        "schemename": "dnet:subject_classification_typologies"
      },
      "value": "CRYSOL"
    },
    {
      "dataInfo": null,
      "qualifier": {
        "classid": "IDPO:00210",
        "classname": "https://disprot.org/assets/data/IDPO_v0.2.owl:00210",
        "schemeid": "dnet:subject_classification_typologies",
        "schemename": "dnet:subject_classification_typologies"
      },
      "value": "ShiftX"
    },
    {
      "dataInfo": null,
      "qualifier": {
        "classid": "IDPO:00216",
        "classname": "https://disprot.org/assets/data/IDPO_v0.2.owl:00216",
        "schemeid": "dnet:subject_classification_typologies",
        "schemename": "dnet:subject_classification_typologies"
      },
      "value": "ENSEMBLE"
    }
  ],
  "embargoenddate": null,
  "lastupdatetimestamp": null,
  "author": null,
  "instance": [
    {
      "refereed": null,
      "hostedby": null,
      "accessright": null,
      "license": null,
      "url": [
        "https://proteinensemble.org/PED00001#P38634_A_1"
      ],
      "measures": null,
      "pid": [
        {
          "dataInfo": {
            "provenanceaction": {
              "classid": "sysimport:actionset",
              "classname": "sysimport:actionset",
              "schemeid": "dnet:provenanceActions",
              "schemename": "dnet:provenanceActions"
            },
            "deletedbyinference": false,
            "inferred": false,
            "inferenceprovenance": null,
            "invisible": false,
            "trust": "0.9"
          },
          "qualifier": {
            "classid": "ped",
            "classname": "ped",
            "schemeid": "dnet:pid_types",
            "schemename": "dnet:pid_types"
          },
          "value": "PED00001#P38634_A_1"
        }
      ],
      "distributionlocation": null,
      "processingchargecurrency": null,
      "alternateIdentifier": [
        {
          "dataInfo": {
            "provenanceaction": {
              "classid": "sysimport:actionset",
              "classname": "sysimport:actionset",
```

```
            "schemeid": "dnet:provenanceActions",
            "schemename": "dnet:provenanceActions"
          },
          "deletedbyinference": false,
          "inferred": false,
          "inferenceprovenance": null,
          "invisible": false,
          "trust": "0.9"
        },
        "qualifier": {
          "classid": "uniprot",
          "classname": "uniprot",
          "schemeid": "dnet:pid_types",
          "schemename": "dnet:pid_types"
        },
        "value": "P38634"
      },
      {
        "dataInfo": {
          "provenanceaction": {
            "classid": "sysimport:actionset",
            "classname": "sysimport:actionset",
            "schemeid": "dnet:provenanceActions",
            "schemename": "dnet:provenanceActions"
          },
          "deletedbyinference": false,
          "inferred": false,
          "inferenceprovenance": null,
          "invisible": false,
          "trust": "0.9"
        },
        "qualifier": {
          "classid": "ped",
          "classname": "ped",
          "schemeid": "dnet:pid_types",
          "schemename": "dnet:pid_types"
        },
        "value": "PED00001#P38634_A_1"
      }
    ],
    "dateofacceptance": null,
    "collectedfrom": {
      "dataInfo": {
        "provenanceaction": {
          "classid": "sysimport:actionset",
          "classname": "sysimport:actionset",
          "schemeid": "dnet:provenanceActions",
          "schemename": "dnet:provenanceActions"
        },
        "deletedbyinference": false,
        "inferred": false,
        "inferenceprovenance": null,
        "invisible": false,
        "trust": "0.9"
      },
      "key": "10|ped_________::pedDatasourceId",
      "value": "Protein Ensemble Database"
    },
    "processingchargeamount": null,
    "instancetype": {
      "classid": "protein",
      "classname": "protein",
      "schemeid": "dnet:publication_resource",
      "schemename": "dnet:publication_resource"
    }
  }
],
"version": null,
"storagedate": null,
"metadataversionnumber": null,
"resulttype": {
  "classid": "dataset",
  "classname": "dataset",
  "schemeid": "dnet:result_typologies",
  "schemename": "dnet:result_typologies"
},
"dateofcollection": "2021-12-10",
"fulltext": null,
"dateoftransformation": null,
"description": [],
"format": null,
"processingchargecurrency": null,
"measures": null,
"dateofacceptance": null,
"coverage": null,
"device": null,
"processingchargeamount": null,
"externalReference": null,
"publisher": null,
"language": null,
"bestaccessright": null,
"country": null,
"extraInfo": null,
```

```
  "originalId": [
    "PED00001#P38634_A_1"
  ],
  "lastmetadataupdate": null,
  "source": null,
  "context": null,
  "title": [
    {
      "dataInfo": {
        "provenanceaction": {
          "classid": "sysimport:actionset",
          "classname": "sysimport:actionset",
          "schemeid": "dnet:provenanceActions",
          "schemename": "dnet:provenanceActions"
        },
        "deletedbyinference": false,
        "inferred": false,
        "inferenceprovenance": null,
        "invisible": false,
        "trust": "0.9"
      },
      "qualifier": {
        "classid": "main title",
        "classname": "main title",
        "schemeid": "dnet:dataCite_title",
        "schemename": "dnet:dataCite_title"
      },
      "value": "PED00001#P38634_A_1 - Structural ensemble of pSic1 (1-90) with phosphorylations at Thr5, Thr33, Thr45, Ser69,
Ser76, Ser80"
    }
  ]
}
```


Relation example:


```
{
  "subRelType": "citation",
  "relClass": "IsCitedBy",
  "dataInfo": {
    "provenanceaction": {
      "classid": "sysimport:actionset",
      "classname": "sysimport:actionset",
      "schemeid": "dnet:provenanceActions",
      "schemename": "dnet:provenanceActions"
    },
    "deletedbyinference": false,
    "inferred": false,
    "inferenceprovenance": null,
    "invisible": false,
    "trust": "0.9"
  },
  "target": "unresolved::20399186::pubmed",
  "lastupdatetimestamp": null,
  "relType": "resultResult",
  "source": "50|ped_________::c7e0527cf8940fca5c78d13c960a6d5c",
  "validationDate": null,
  "collectedfrom": [
    {
      "dataInfo": {
        "provenanceaction": {
          "classid": "sysimport:actionset",
          "classname": "sysimport:actionset",
          "schemeid": "dnet:provenanceActions",
          "schemename": "dnet:provenanceActions"
        },
        "deletedbyinference": false,
        "inferred": false,
        "inferenceprovenance": null,
        "invisible": false,
        "trust": "0.9"
      },
      "key": "10|ped_________::pedDatasourceId",
      "value": "Protein Ensemble Database"
    }
  ],
  "validated": false,
  "properties": [
    {
      "dataInfo": null,
      "key": "RelationDate",
      "value": null
    }
  ]
}
```

Conclusion and Future work

In this report we described an extended hadoop-based aggregator for Bioschema data sources. We applied the extended hadoop-based aggregator to the Protein profile (https://bioschemas.org/profiles/Protein/0.11-RELEASE). Currently there are 9 sites that provide data according to Protein profile (https://bioschemas.org/liveDeploys#nav-profile).
In our future work we want to enhance the proposed solution in order to harvest the data related to the following Bioschema profiles:

- Gene (https://bioschemas.org/profiles/Gene/1.0-RELEASE), 7 available sites.
- MolecularEntity (https://bioschemas.org/profiles/MolecularEntity/0.5-RELEASE) , 6 available sites.
- ComputationalTool (https://bioschemas.org/profiles/ComputationalTool/1.0-RELEASE) , 8 available sites.

We also plan to evaluate the development of a web app with BMUSE embedded that can handle a scheduled scraping process. With such a service available we could avoid using BMUSE tool inside hadoop cluster.

Acknowledgements

We would like to thank Andreas Czerniak (University of Bielefeld, Germany) for his contribution in data mapping from Bioschemas to OpenAIRE fields.