



Bioschemas:

schema.org for Life Sciences

Leyla Jael Garcia Castro (ljgarcia@ebi.ac.uk)
Protein Function Development team, EMBL-EBI

Elixir

Hinxton, 15th of March 2017



Bioschemas

Community initiative built on top of schema.org, aiming to improve data discoverability and interoperability in Life Sciences



Creating Life Sciences schemas/specifications



Testing adoption in Life Sciences key resources



Evaluating pros and cons



Implementing practical examples → proof of concept application





Use cases



Findability

Easily finding proteins, samples, phenotypes and so on resources on the web



Resource index

Getting a quick view on how ontologies are used in resources marked with Bioschemas



Accessibility

Gathering structured information from different life sciences resources without dealing with multiple formats



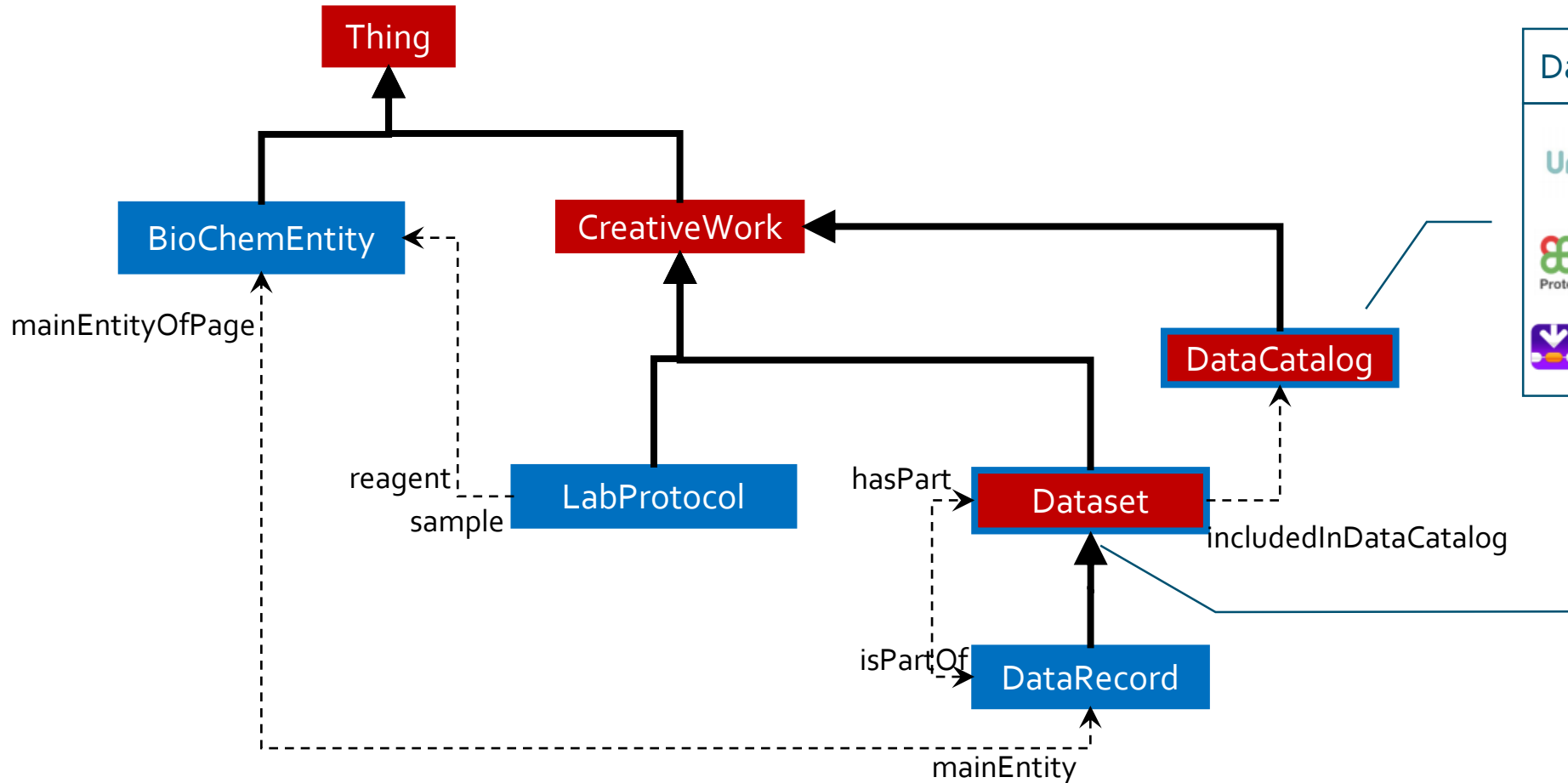
Summarization

Getting a quick summary with information from different resources as well as links to them





Specifications



Data repositories, e.g.,



BioSamples



Datasets and releases, e.g.,

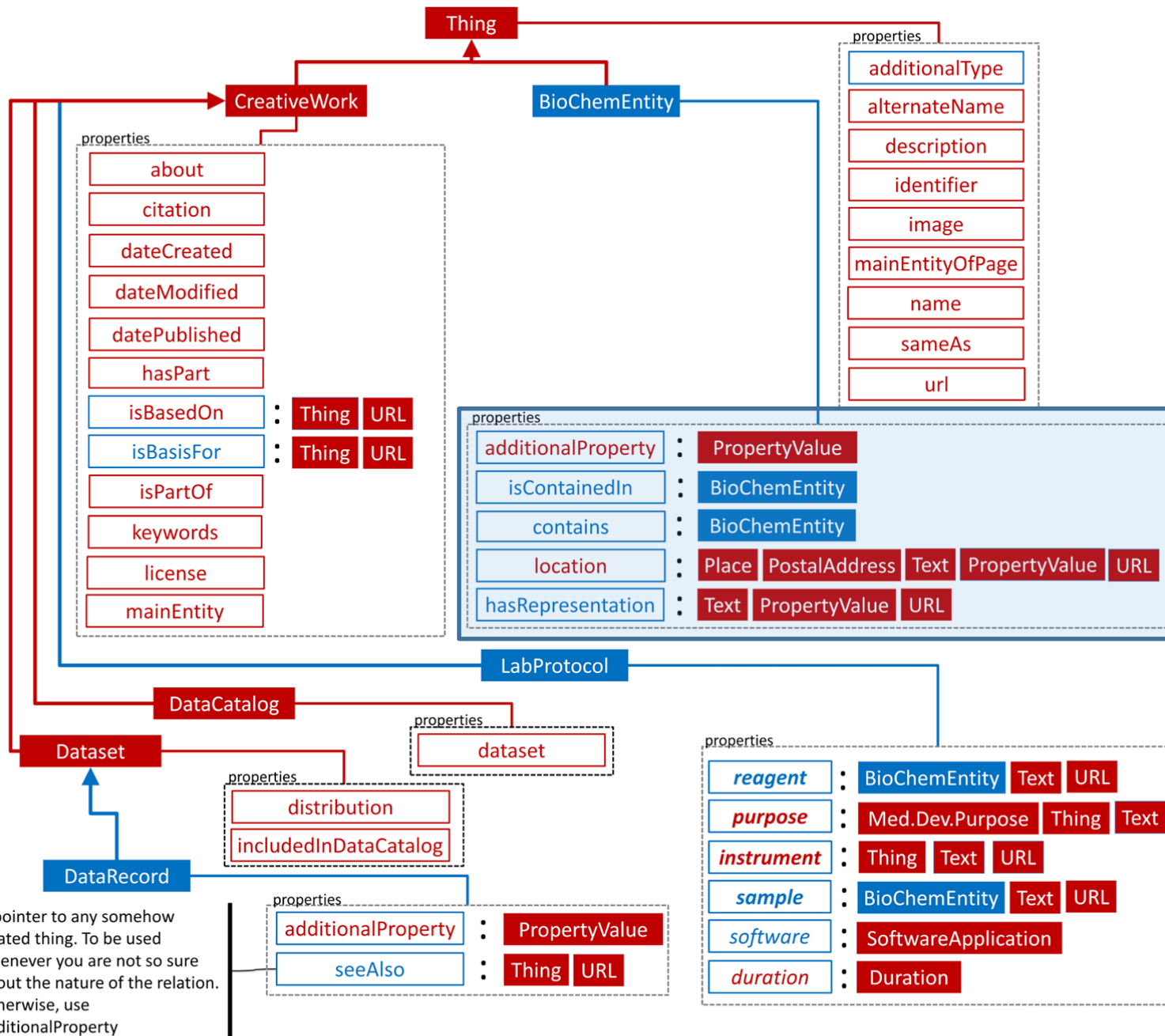
- UniProtKB 2017_07
- InterPro 64.0
- WTCCC1 project Hypertension (HT) samples





A more detailed overview

- BioChemEntity
 - Flexible and extensible wrapper
 - Additional properties
 - 1st option → reuse external well known ontology terms
 - 2nd option → use schema:additionalProperty to mint your own property
- DataRecord
- LabProtocol
 - Based on SMARTProtocols
 - Specific parts via schema:hasPart → objectives, limitations, etc.



Minimum: bold and italic
Recommended: italic
Optional: Regular font



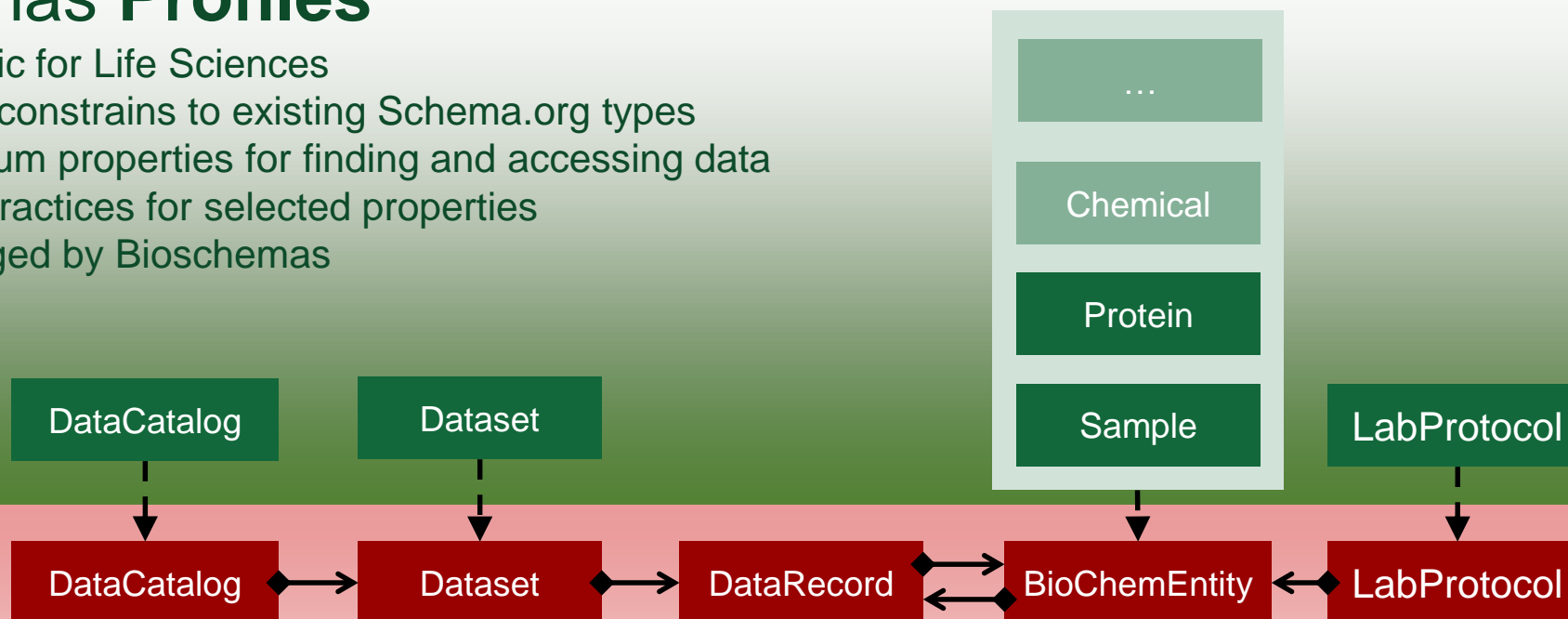
schema.org



Types and profiles

Bioschemas Profiles

- Specific for Life Sciences
- Apply constraints to existing Schema.org types
- Minimum properties for finding and accessing data
- Best practices for selected properties
- Managed by Bioschemas



Schema.org Types

- Generic data model
- Generous list of properties to describe data types
- Managed by Schema.org





UniProt, a case study

UniProt

UniProtKB

Advanced Search

BLAST Align Retrieve/ID mapping Peptide search

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

UniProtKB
UniProt Knowledgebase

Swiss-Prot (555,426)
Manually annotated and reviewed.

TrEMBL (89,396,316)
Automatically annotated and not reviewed.

UniRef
Sequence clusters

UniParc
Sequence archive

Proteomes

Supporting data

Literature citations
Cross-ref. databases

Taxonomy
Diseases
XXX

Subcellular locations
Keywords

News

Forthcoming changes
Planned changes for UniProt

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

UniProt release 2017_07
A pseudogene turns into an active DNA methyltransferase dedicated to male fertility

News archive

Getting started

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

BLAST
Find regions of similarity between your sequences

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

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

UniProt data

Download latest release
Get the UniProt data

Statistics
View Swiss-Prot and TrEMBL statistics

How to cite us
The UniProt Consortium

Submit your data
Submit your sequences and annotation updates

SPARQL
Query UniProt data using a SQL like graph query

Protein spotlight

A Taste Of Light
August 2017

Light gave life a chance to be. Without it, our planet would not be inhabited by so many living beings of all shapes and sizes. Over time, animals, plants and all sorts of microorganisms have emerged and evolved using this source of photons in different ways. Like hosts of other organisms, we owe light for various reasons that we can discuss in individual entities that make up our environment, as well as movement within it...

Data repository
for protein
sequences and
annotations

Multiple datasets
with releases
every 4 weeks

Downloads



Protein profile, based on BioChemEntity

UniProtKB - P00519 (ABL1_HUMAN)

Protein Identifier

Basket

Display

Entry

Publications

Feature viewer

Feature table

All None

☒ Function

☒ Names & Taxonomy

☒ Subcellular location

☒ Pathology & Biotech

☒ PTM / Processing

☒ Expression

☒ Interaction

☒ Structure

☒ Family & Domains

☒ Sequences (2)

☒ Similar proteins

☐ Cross-references

☒ Entry information

☒ Miscellaneous

BLAST Align Format Add to basket History

Feedback Help video Other tutorials and videos

Protein Tyrosine-protein kinase ABL1

Gene ABL1

Organism Homo sapiens (Human)

Status Reviewed - Annotation score: ●●●●● - Experimental evidence at protein levelⁱ

Transcribed from Gene

Organism

Functionⁱ

Non-receptor tyrosine-protein kinase that plays a role in many key processes linked to cell growth and survival such as cytoskeleton remodeling in response to extracellular stimuli, cell motility and adhesion, receptor endocytosis, autophagy, DNA damage response and apoptosis. Coordinates actin remodeling through tyrosine phosphorylation of proteins controlling cytoskeleton dynamics like WASF3 (involved in branch formation); ANXA1 (involved in membrane anchoring); DBN1, DBNL, CTTN, RAPH1 and ENAH (involved in signaling); or MAPT and PXN (microtubule-binding proteins). Phosphorylation of WASF3 is critical for the stimulation of lamellipodia formation and cell migration. Involved in the regulation of several biological processes through phosphorylation of key regulators of these processes such as BCAR1, CRK, CRKL, DOK1, EFS or NEDD9. Phosphorylates multiple receptor tyrosine kinases and more particularly promotes endocytosis of EGFR, facilitates the formation of neuromuscular synapses through MUSK, inhibits PDGFRB-mediated chemotaxis and modulates the endocytosis of activated B-cell receptor complexes. Other substrates which are involved in endocytosis regulation are the caveolin (CAV1) and RIN1. Moreover, ABL1 regulates the CBL family of ubiquitin ligases that drive receptor down-regulation and actin remodeling. Phosphorylation of CBL leads to increased EGFR stability. Involved in late-stage autophagy by regulating positively the trafficking and function of lysosomal components. ABL1 targets to mitochondria in response to oxidative stress and thereby mediates mitochondrial dysfunction and cell death. In response to oxidative stress, phosphorylates serine/threonine kinase PRKD2 at 'Tyr-717' (PubMed:28428613). ABL1 is also translocated in the nucleus where it has DNA-binding activity and is involved in DNA-damage response and apoptosis. Many substrates are known mediators of DNA repair: DDB1, DDB2, ERCC3, ERCC6, RAD9A, RAD51, RAD52 or WRN. Activates the proapoptotic pathway when the DNA damage is too severe to be repaired. Phosphorylates TP73, a primary regulator for this type of damage-induced apoptosis. Phosphorylates the caspase CASP9 on 'Tyr-153' and regulates its processing in the apoptotic response to DNA damage. Phosphorylates PSMA7 that leads to an inhibition of proteasomal activity and cell cycle transition blocks. ABL1 acts also as a regulator of multiple pathological signaling cascades during infection. Several known tyrosine-phosphorylated microbial proteins have been identified as ABL1 substrates. This is the case of A36R of Vaccinia virus, Tir (translocated intimin receptor) of pathogenic E.coli and possibly Citrobacter, CagA (cytotoxin-associated gene A) of H.pylori, or AnkA (ankyrin repeat-containing protein A) of A.phagocytophilum. Pathogens can hijack ABL1 kinase signaling to reorganize the host actin cytoskeleton for multiple purposes, like facilitating intracellular movement and host cell exit.

Protein name

Function

Citations

Associated to diseases





Data Catalog

Describes metadata for data repositories and data catalogues so they can be more easily indexed by search engines and registries.

Specification

<http://bioschemas.org/specifications>

Based on schema.org/DataCatalog

Property	Expected Type	Description	CD
keywords	Text	Keywords or tags used to describe this content. Multiple entries in a keywords list are typically delimited by commas.	MANY
provider	Organization Person	The service provider, service operator, or service performer; the goods producer. Another party (a seller) may offer those services or goods on behalf of the provider. A provider may also serve as the seller.	ONE
description	Text	A description of the item.	ONE
name	Text	The name of the item.	ONE
url	URL	URL of the item.	ONE
dataset	Dataset	A dataset contained in this catalog.	MANY
citation	CreativeWork Text	A citation or reference to another creative work, such as another publication, web page, scholarly article, etc.	MANY
dateModified	Date DateTime	The date on which the CreativeWork was most recently modified or when the item's entry was modified within a DataFeed.	ONE
license	CreativeWork URL	A license document that applies to this content, typically indicated by URL.	ONE
publication	PublicationEvent	A publication event associated with the item.	MANY
sourceOrganization	Organization	The Organization on whose behalf the creator was working.	MANY
alternateName	Text	An alias for the item.	MANY
identifier	PropertyValue Text URL	The identifier property represents any kind of identifier for any kind of Thing , such as ISBNs, GTIN codes, UUIDs etc.	MANY



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



Data catalog →
minimum

```
  "name": "UniProt",  
  "description": "The Universal Protein Resource (UniProt) is a  
comprehensive resource for protein sequence and annotation  
data",  
  "url": "http://www.uniprot.org",  
  "keywords": "protein, protein sequence, protein annotation",  
  "provider": {  
    "@type": "Organization",  
    "name": "UniProt Consortium"  
  }  
}
```



Datasets

Describes metadata for datasets so they can be more easily indexed by search engines and registries.

Specification

<http://bioschemas.org/specifications>

Based on schema.org/Dataset

Property	Expected Type	Description	CN
name	Text	<i>The name of the item.</i> It is a descriptive name of the dataset	One
description	Text	<i>A description of the item.</i> It is a short summary describing a dataset.	One
url	URL	<i>The URL of the item.</i> It is the location of a page describing the dataset.	One
identifier	PropertyValue or Text or URL	<i>The identifier property represents any kind of identifier for any kind of Thing, such as ISBNs, GTIN codes, UUIDs etc.</i>	Many
keywords	Text	<i>Keywords or tags used to describe this content. Multiple entries in a keywords list are typically delimited by commas. These keywords provide a summary of the dataset.</i>	Many
includedInDataCatalog	DataCatalog	<i>A data catalog which contains this dataset.</i>	Many
creator	Text	<i>The creator/author of this CreativeWork. This is the same as the Author property for CreativeWork.</i> The name of the dataset creator (person or organization)	Many
version	Text , Number	The version number for this dataset	One
variableMeasured	Text , PropertyValue	What does the dataset measure? (e.g., temperature, pressure)	Many
measurementTechnique	Text	A technique or technology used for measuring the corresponding variable(s) (described using variablesMeasured)	Many
citation	Text	A citation for a publication that describes the dataset	Many
license	CreativeWork , URL	A license under which the dataset is distributed	Many
distribution	DataDownload	A downloadable form of this dataset, at a specific location, in a specific format	Many





Datasets → minimum

```
{  
  "@type": "Dataset",  
  "@id": "http://www.uniprot.org/uniprot/",  
  "name": "UniProt Knowledgebase (UniProtKB)",  
  "description": "The UniProt Knowledgebase (UniProtKB) ...",  
  "url": "http://www.uniprot.org/uniprot/",  
  "identifier": "UniProtKB",  
  "keywords": "protein, protein sequence, protein annotations,  
knowledgebase, TrEMBL, Swiss-Prot"  
}
```



DataRecord

Describes a data record included in a dataset, so they can be more easily indexed by search engines and registries.

- mainEntity → link to BioChemEntity
- isPartOf → link to dataset
- url → link to official webpage

Extending schema.org/Dataset

Property	Expected type	Description	CN
identifier	PropertyValue or Text or URL	The identifier property represents any kind of identifier for any kind of Thing , such as ISBNs, GTIN codes, UUIDs etc.	One
mainEntity	Thing	Indicates the primary entity described in some page or other CreativeWork.	Many
isPartOf	CreativeWork	Indicates a CreativeWork that this CreativeWork is (in some sense) part of.	Many
url	URL	URL of the item.	Many
seeAlso	URL	Link to other related data records	Many





Protein record example → minimum & recommended

```
{  
  "@context": "http://schema.org",  
  "@type": "Record",  
  "@id": "http://www.identifiers.org/uniprot/P00519",  
  
  "identifier": "P00519",  
  "mainEntity": { ... },  
  "isPartOf": {  
    "@type": "Dataset",  
    "@id": "http://www.uniprot.org/news/2017/03/15/release"  
  },  
  
  "additionalType": "http://purl.uniprot.org/core/Protein",  
  "url": "http://www.uniprot.org/uniprot/P00519",  
}
```





Protein record example → optional

```
"sameAs": "http://purl.uniprot.org/uniprot/P00519",

"citation": [
  {
    "@id": "http://www.identifiers.org/pubmed/10194451",
    "@type": "ScholarlyArticle",
    "name": { "@language": "en", "@value": "A novel SH2-containing ..." },
    "sameAs": [
      "https://www.ncbi.nlm.nih.gov/pubmed/10194451", "http://europepmc.org/abstract/MED/10194451",
      "http://purl.uniprot.org/citations/10194451"
    ]
  },
  {
    "@id": "http://www.identifiers.org/pubmed/9037071",
    "@type": "ScholarlyArticle",
    "name": { "@language": "en", "@value": "Regulation of DNA ..." },
    "sameAs": [
      "https://www.ncbi.nlm.nih.gov/pubmed/9037071", "http://europepmc.org/abstract/MED/9037071",
      "http://purl.uniprot.org/citations/10194451"
    ]
  }
],
"dateCreated": "1986-07-21",
"dateModified": "2017-03-15",
"distribution": {
  "@type": "DataDownload",
  "url": "http://www.uniprot.org/uniprot/P00519.fasta"
},

"seeAlso": [
  "http://www.identifiers.org/pdb/1AB2", "http://www.identifiers.org/pdb/1ABL"
]
```




BioChemEntity

Protein profile → Describes a protein

- Protein type → http://purl.obolibrary.org/obo/PR_00000001
- Additional properties → reuse well know ontology terms
 - Some of them minimum or recommended
 - Any other is optional

Extending schema.org/Thing

Property	Expected type	Description	CN
identifier	PropertyValue or Text or URL	The identifier property represents any kind of identifier for any kind of Thing , such as ISBNs, GTIN codes, UUIDs etc.	ONE
isContainedIn	BioChemEntity	Indicates a BioChemEntity that this PhysicalEntity is (in some sense) part of	MANY
alternateName	Text	An alias for the item.	MANY
description	Text	A description of the item.	ONE
name	Text	The name of the item.	ONE
url	URL	URL of the item.	ONE
SIO:is-related-to	URL, MedicalCondition	Disease associated to this protein	MANY
SIO:is-transcribed-from	URL, BioChemEntity	Gene which this protein was transcribed from	MANY





Protein example → minimum & recommended

```
{
  "@context": [
    "http://schema.org", {"@base": "http://schema.org"},
    {
      "Gene": { "@id": "http://purl.obolibrary.org/obo/SO_0000704" },
      "Protein": { "@id": "http://purl.obolibrary.org/obo/PR_000000001" },
      "transcribedFrom": { "@id": "http://semanticscience.org/resource/SIO_010081" },
      "associatedTo": { "@id": "http://semanticscience.org/resource/SIO_000001" }
    }
  ],

```

```
"@type": ["BioChemEntity", "Protein"],
```

```
"identifier": "P00519",
```

```
"additionalType": "http://semanticscience.org/resource/SIO_010043",
"alternateName": ["ABL", "JTK7"],
"description": "Non-receptor tyrosine-protein kinase that plays a role...",
"name": "ABL1",
"url": "http://www.uniprot.org/uniprot/P00519",
```

Context providing IRIs for
types and additional
properties

Community votes to reach
agreement



Protein example → minimum & recommended

```
"isContainedIn": {  
  "@type": "BioChemEntity",  
  "additionalType": "http://purl.obolibrary.org/obo/OBI_0100026",  
  "identifier": "9606", "name": "Homo sapiens",  
  "url": "http://purl.bioontology.org/ontology/NCBITAXON/9606",  
  "sameAs": "http://purl.uniprot.org/taxonomy/9606"  
},  
"associatedTo": {  
  "@type": "MedicalCondition",  
  "additionalType": "http://semanticscience.org/resource/SIO_010299",  
  "name": "Leukemia, chronic myeloid (CML)",  
  "code": { "@type": "MedicalCode", "code": "608232", "codingSystem": "OMIM" },  
  "sameAs": "http://www.uniprot.org/diseases/DI-03735"  
},  
"transcribedFrom": {  
  "@type": ["BioChemEntity", "Gene"],  
  "additionalType": "http://purl.obolibrary.org/obo/SO_0000704",  
  "identifier": "ABL1", "name": "ABL1"  
},  
}
```

```
'http://semanticscience.org/resource/SIO_000095': 'http://pfam.xfam.org/clan/CL0001',  
}
```





From specifications to adoption, from adoption to benefits

- Current status → reaching agreement on ontology terms for profile types and additional properties
 - e.g., protein type, transcribed from gene or disease association
- Ongoing → Adoption by key resources
 - e.g., UniProt, InterPro, Protein Data Bank
- Next step → proof of concept → summaries

Summary for a protein search



Insulin

Insulin decreases blood glucose concentration. It increases cell permeability to monosaccharides, amino acids and fatty acids... [More](#)

Data providers

1	UniProt	P01308
2	Ensemble	ENSG000...
3	RefSeq (protein)	NP_000196

Literature

All Articles (22493411)  Europe PMC [PubMed](#)

Reviews (1746164)  Europe PMC [PubMed](#)

Free Full Text Articles (3545104)  Europe PMC [PubMed](#)

People also search for



Thank you



<http://bioschemas.org/>

<http://bioschemas.org/howtojoin/>