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

Accessibility

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

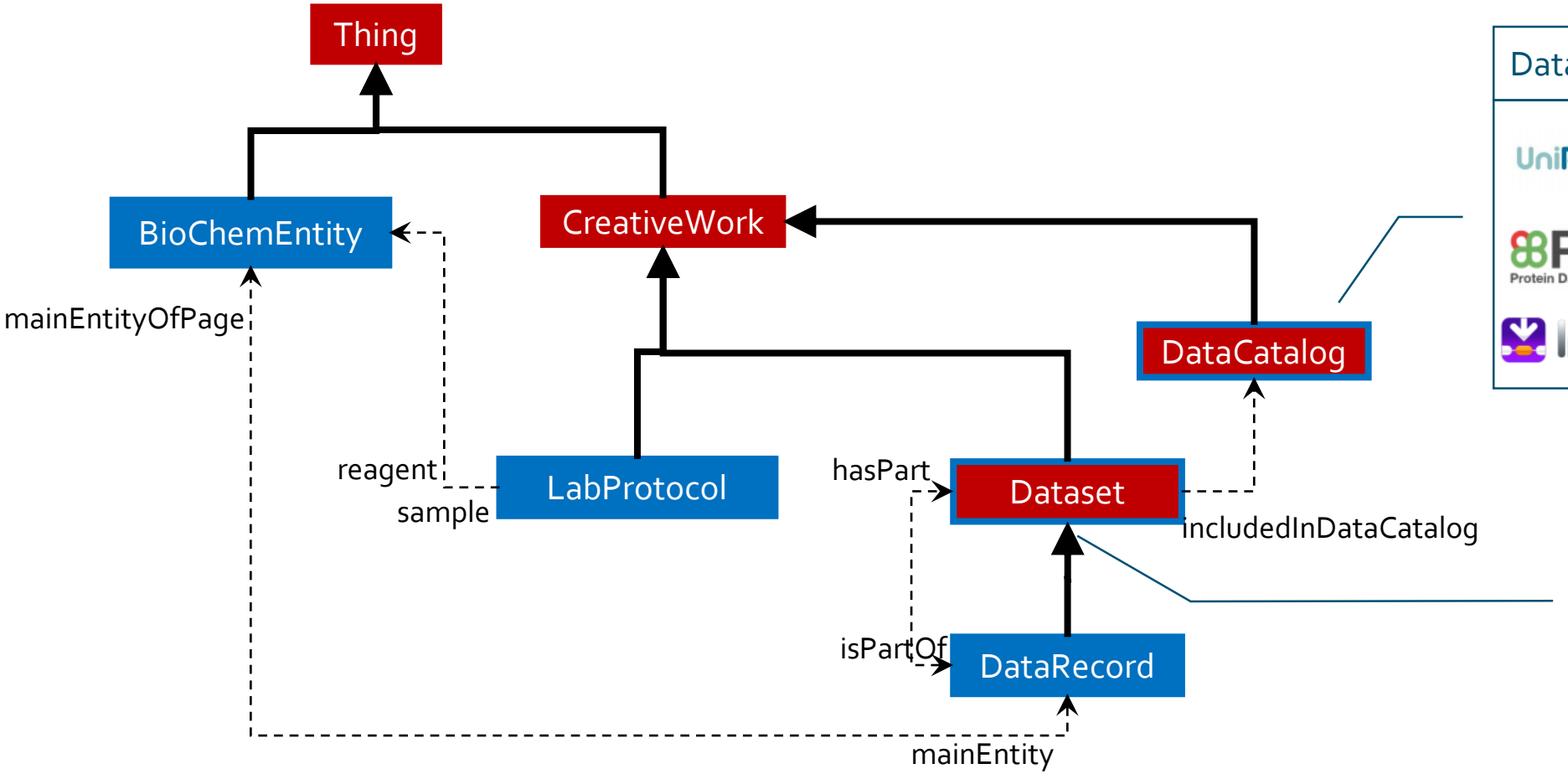
Resource index

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

Summarization

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

Specifications

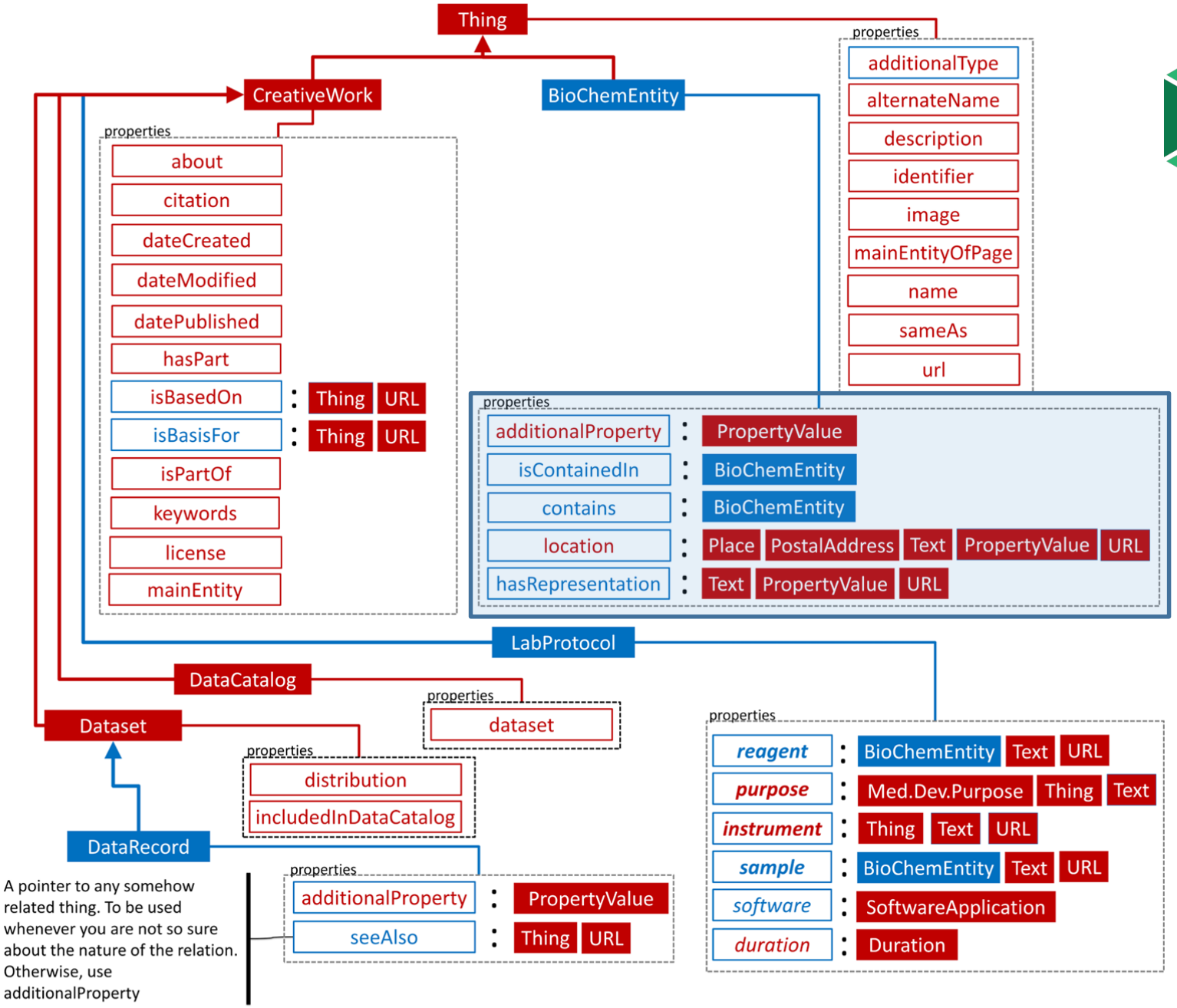


Data repositories, e.g.,

UniProt BioSamples PDBe Protein Data Bank in Europe European Genome-phenome Archive InterPro

Datasets and releases, e.g.,

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



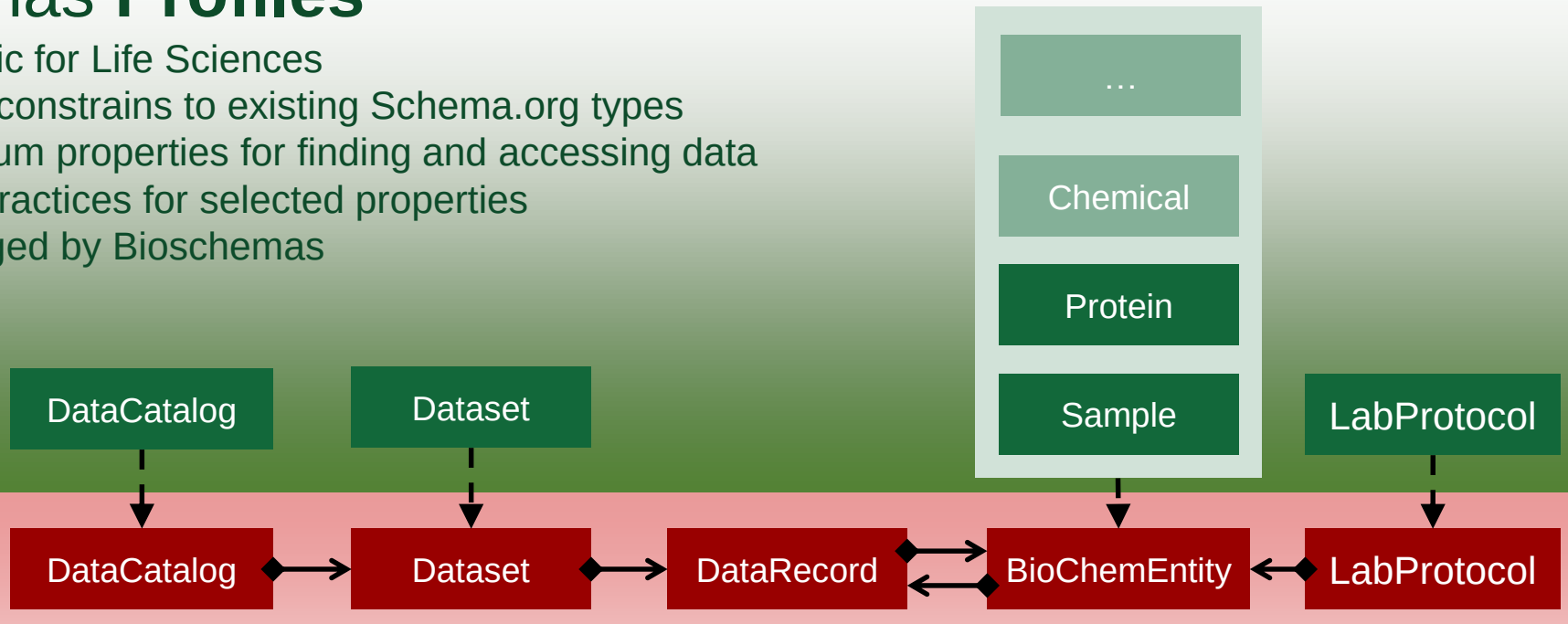
A more detailed overview

- 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.

Types and profiles

Bioschemas **Profiles**

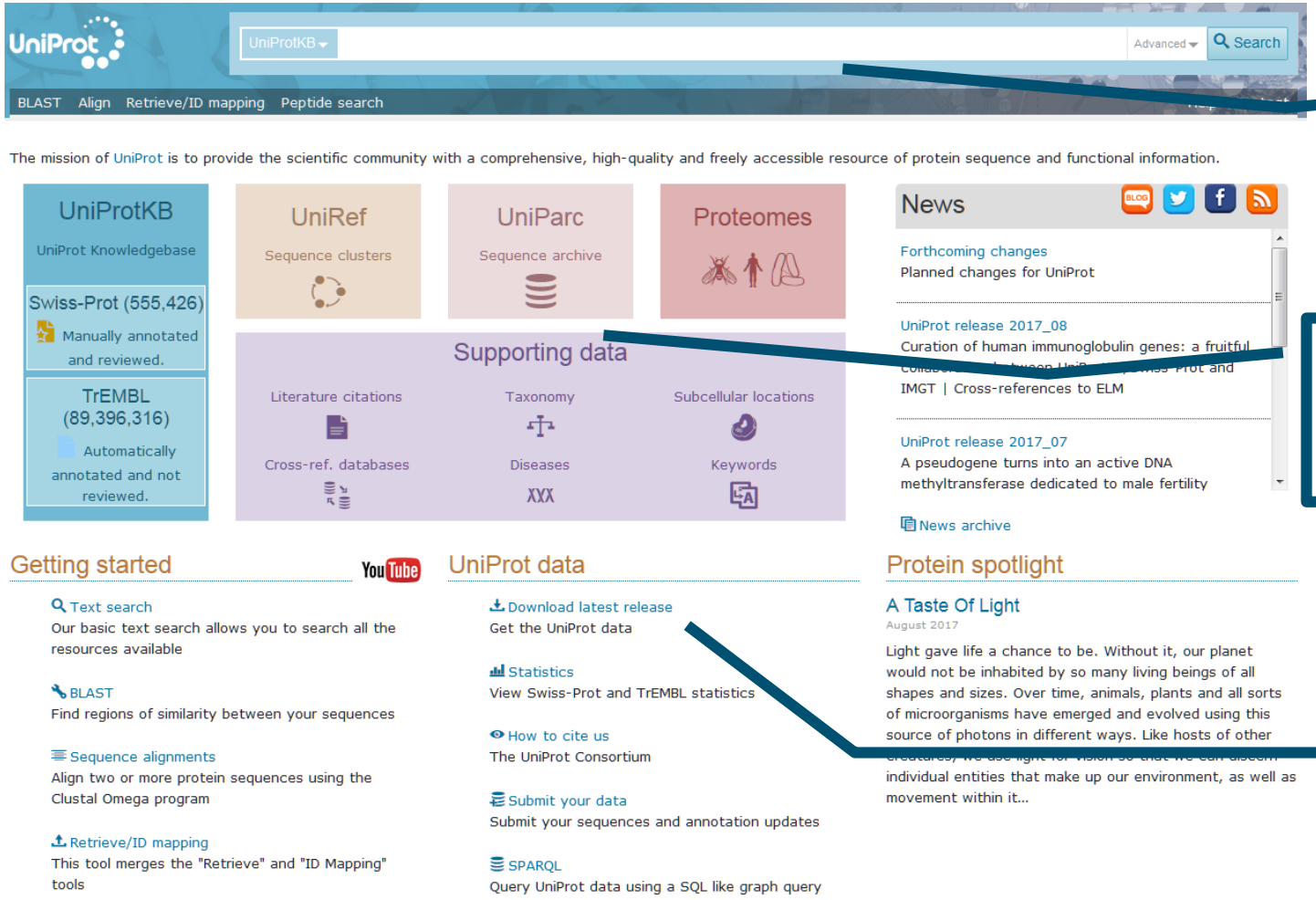
- Specific for Life Sciences
- Apply constrains to existing Schema.org types
- Minimum properties for finding and accessing data
- Best practices for selected properties
- Managed by Bioschemas



Schema.org **Types**

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

UniProt, a case study

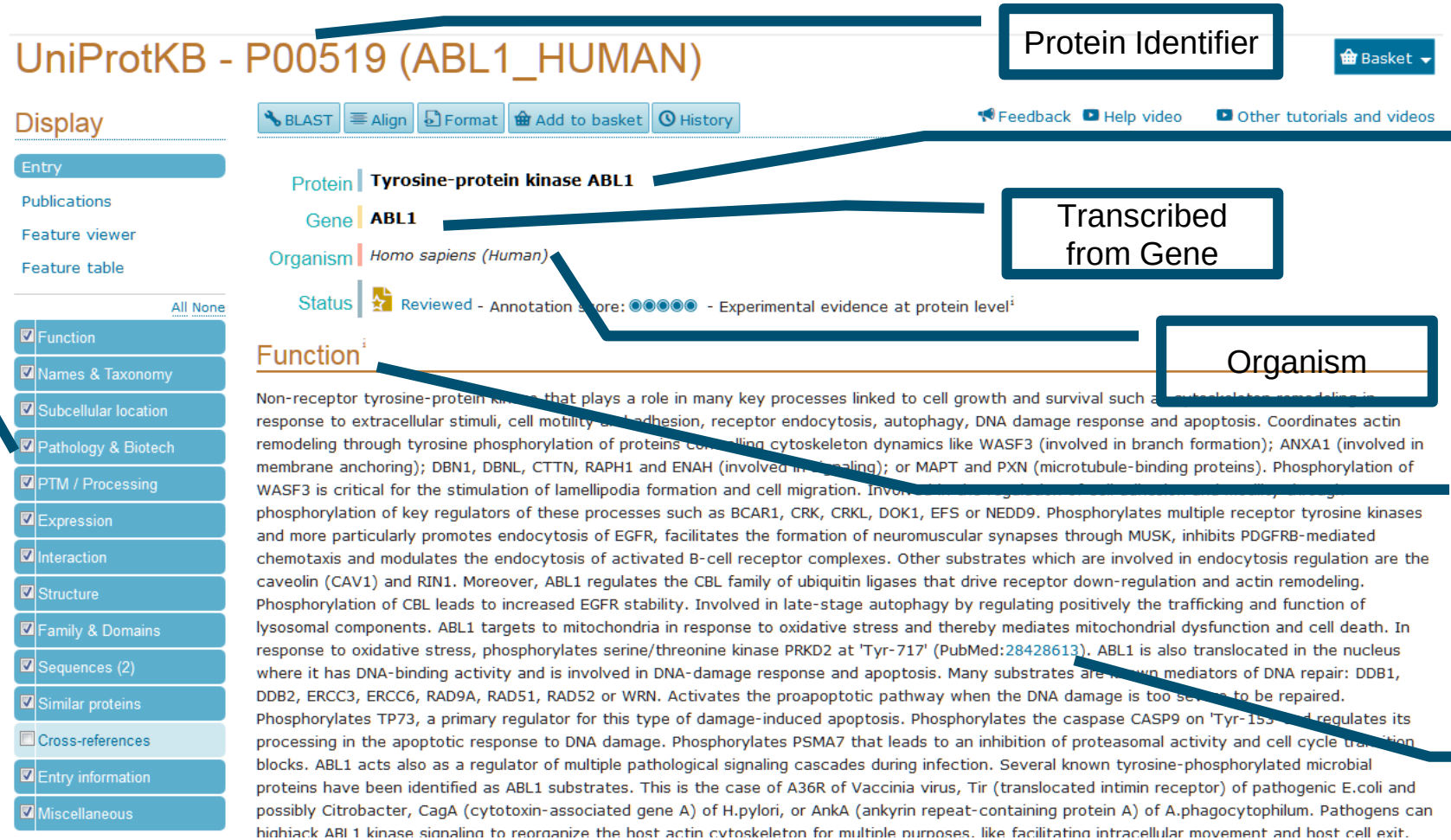


Data repository for protein sequences and annotations

Multiple datasets with releases every 4 weeks

Downloads

Protein profile, based on BioChemEntity



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

Data catalog → minimum

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

 "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	*The name of the item.* It is a descriptive name of the dataset	*One*
description	Text	*A description of the item.* It is a short summary describing a dataset.	*One*
url	URL	*The URL of the item.* It is the location of a page describing the dataset.	*One*
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.*	*Many*
keywords	Text	*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.	*Many*
includedInDataCatalog	DataCatalog	*A data catalog which contains this dataset.*	Many
creator	Text	*The creator/author of this CreativeWork. This is the same as the Author property for CreativeWork.* 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_000000001
- 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

Thank you

http://bioschemas.org/
http://bioschemas.org/howtojoin/