Persistent Identifiers in the Life Sciences

Florian Gräf

Europe PMC

The European Molecular Biology Laboratory

80+ nationalities

>1600 personnel

6 sites in Europe

Heidelberg, Germany

Main Laboratory

Tissue Biology, Disease Modeling

Barcelona, Spain

Hinxton, Cambridge, UK

Bioinformatics

Mouse Biology

Monterotondo, Rome, Italy

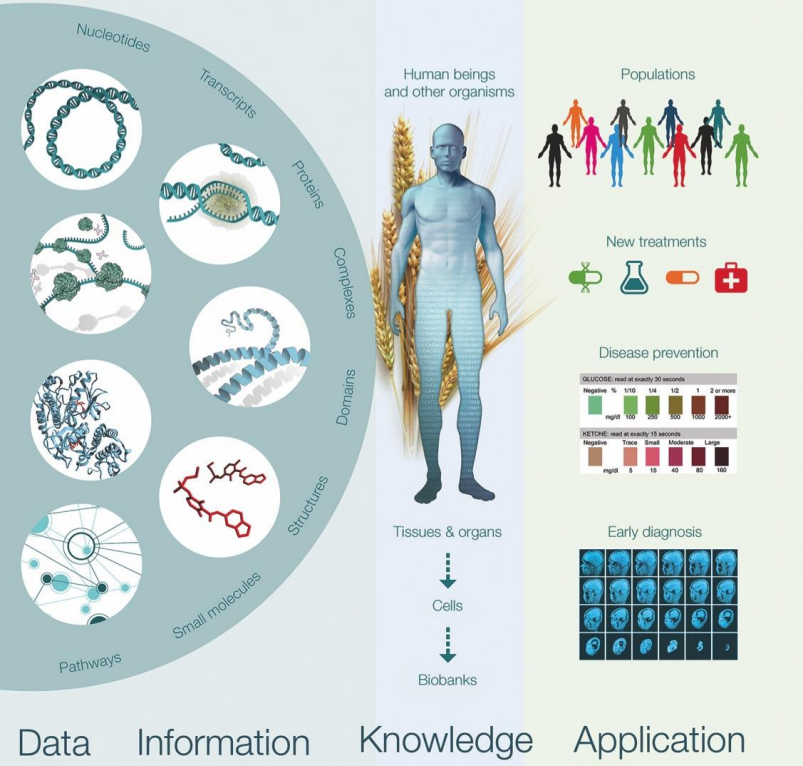
Grenoble, France

Structural Biology

Structural Biology

Hamburg, Germany

From molecules to medicine



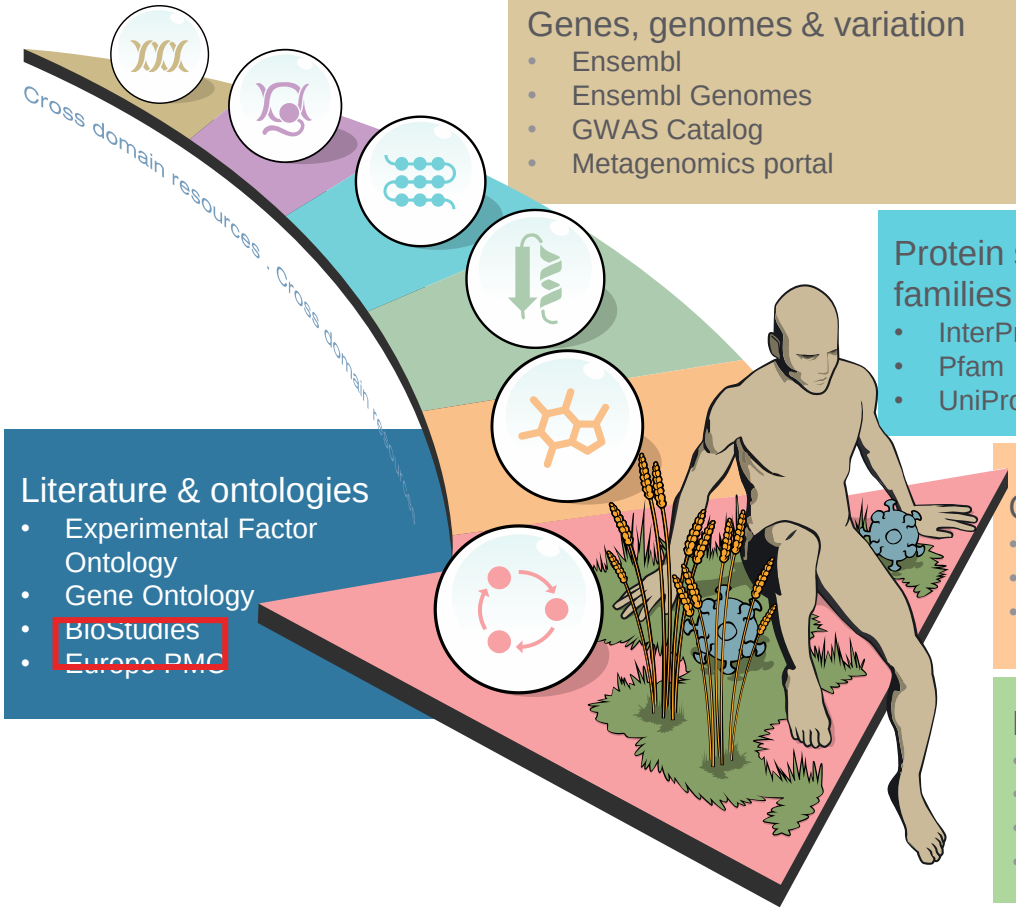
We are always seeking new ways to read and understand DNA

New technologies provide ways to collect, compare and visualize molecular information

Bioinformatics enables new applications:

- molecular medicine
- agriculture
- food
- environmental sciences

Data resources at EMBL-EBI



Genes, genomes & variation
- Ensembl
- Ensembl Genomes
- GWAS Catalog
- Metagenomics portal

Gene, protein & metabolite expression
- Expression Atlas
- Metabolights
- PRIDE
- RNA Central

Protein sequences, families & motifs
- InterPro
- Pfam
- UniProt

Molecular structures
- Protein Data Bank in Europe
- Electron Microscopy Data Bank

Literature & ontologies
- Experimental Factor Ontology
- Gene Ontology
- BioStudies
- Europe PMC

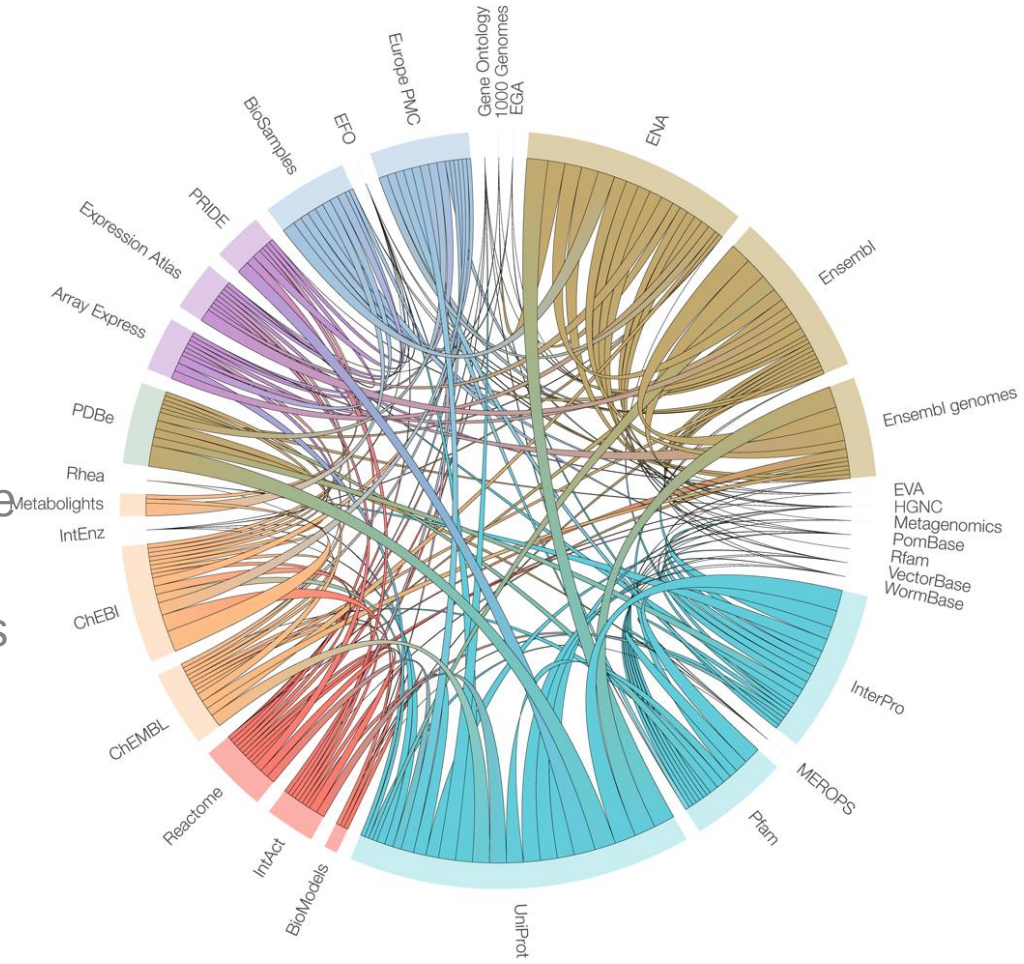
Chemical biology
- ChEBI
- ChEMBL
- SureChEMBL

Systems
- BioModels
- BioSamples
- Enzyme Portal
- IntAct
- Reactome

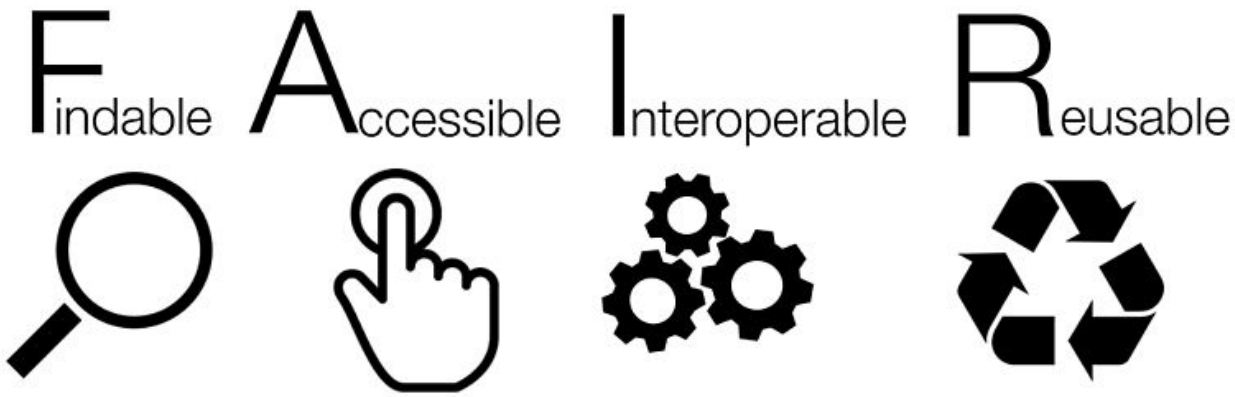
Molecular Archives
- European Nucleotide Archive
- European Variation Archive
- European Genome-phenome Archive
- ArrayExpress

Database interactions

- Our collaborative community facilitates social, scientific and technical interactions
- This image shows internal interactions between data resources, as determined by the exchange of data.
- The width of each internal arc is weighted according to the number of different data types exchanged.



FAIR data



Big data, big demand

~27 million
requests to EMBL-EBI websites every day

Scientists at over
3.2 million
unique IP addresses use EMBL-EBI websites

EMBL-EBI delivered
152 million
jobs to its users in 2016

120 petabytes
of storage capacity in our data centres

Generic vs structured/specific data archives

Data submission

Use this data submission wizard to find the right archive for your data in a few simple steps.

1 What **type of data** do you have?

DNA/RNA sequence | Expression data | Protein data | Structures | Systems | Chemical biology | Ontologies

Multi-omics or other cross-domain study

Why submit data to an archive?

Submission of primary data and derived information to public data repositories is an essential step in the scientific process. Through submission, the scientific community is fed the raw materials for the building and maintenance of the complete and up-to-date data sets that support searches and analysis on the latest sequences, structures and molecular profiles of living systems. Serving as a complement to the literature publication process and supporting early data sharing, the EBI offers a number of submission services appropriate for different types and scales of data.

DRYAD

figshare

zenodo

BioStudies.

Biological metadata – critical to scientific reuse

Specific: organism, tissue, phenotype, location, process …

deep search
analysis
computation

```
FT   source          1..315242
FT                   /organism="Homo sapiens"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:9606"
FT   mRNA            join(133806..133969,145842..146028,151911..152061,
FT                   162216..162363,162790..162956,164592..164729,
FT                   167866..167999,178281..178360,183171..183244,
FT                   190181..191728)
FT                   /gene="RHD"
FT   exon            133806..133969
FT                   /gene="RHD"
FT                   /number=1
FT   CDS             join(133822..133969,145842..146028,151911..152061,
```

BioStudies.

Generic: title, submitters, date, file format, versi

figshare DRYAD zenodo

citation
basic search

Wagner F.F., 23-APR-2002, *TPA: Homo sapiens SMP1 gene, RHD gene and RHCE gene*, INSDC, 14-NOV-2006 (Rel. 89, Last updated, Version 7). BN000065

Persistent Identifiers in the Life Sciences

- Accession numbers
  - heterogeneous, recognizable, community accepted
  - 4XNR, 9606, JF803844, ENSG00000139618, P13569
- DOIs
  - Journals and generic data resources (FigShare, Dryad, Zenodo)
  - Occasional use for high-level datasets as 2◦ PID
- ORCIDs
  - 4.5M articles, 0.5M published authors in Europe PMC
  - Application to submitted datasets

Data collections and identifiers in life sciences

- Actionable identifiers embedded in URLs

https://www.ebi.ac.uk/pdbe/entry/pdb/2gc4

http://www.wormbase.org/db/gene/gene?name=WBGene00000001;class=Gene

http://www.ebi.ac.uk/ena/data/view/Taxon:9606

- The same data collection is often provided by many alternative physical locations

http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=9606

http://www.ebi.ac.uk/ena/data/view/Taxon:9606

Challenges

- Multiple URLs for the same collection make object unification challenging
- Which resource should be used for annotation or citation
- A given location may be down, or change its URL, resulting in dead links

Identifiers.org

- *Identifiers.org* system provides unique stable, resolvable and location-independent URIs to identify and locate life science data
- Promotes *Findable*, *Accessible*, *Interoperable* and *Re-usable* (FAIR) data
- Over 10 years supporting data integration
- Community driven
- Free to use

Compact Identifiers

- A unique prefix indicating the assigning authority
- A locally assigned database identifier sometimes called an accession
- An additional provider level prefix (provider_code) to identify individual hosts

prefix:accession

provider code/prefix:accession

Europe PMC

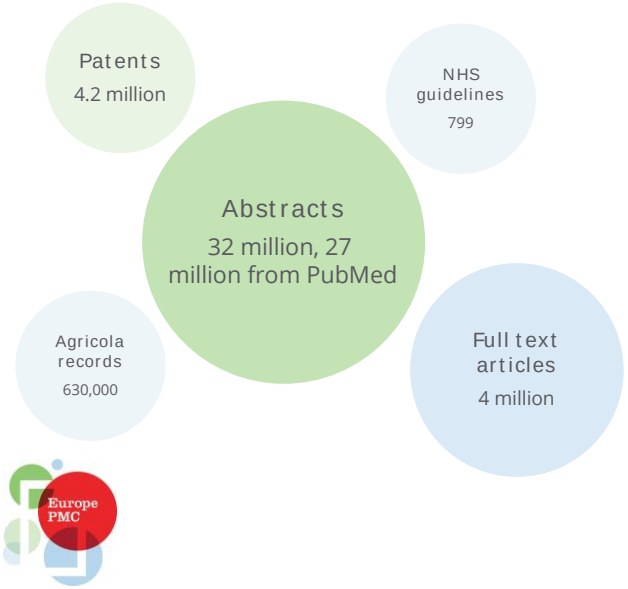
- Since 2007 (UKPMC)
- 30+ million documents of which 4 million full text, >25 million PubMed abstracts
- **Enrichments:** ORCIDs, citations, named entities, DOIs, data links
- Website, web services, FTP

- A PMCI partner: PMC & PMC Canada

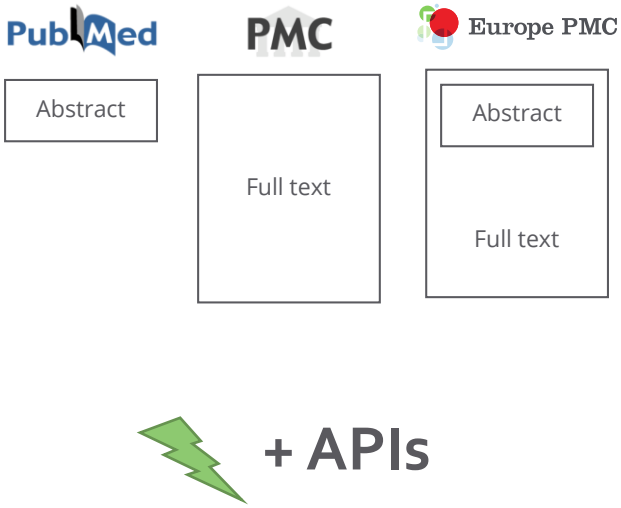
Content in Europe PMC

- Europe PMC is a partner in PubMed Central International.
  - Content is freely shared between the nodes

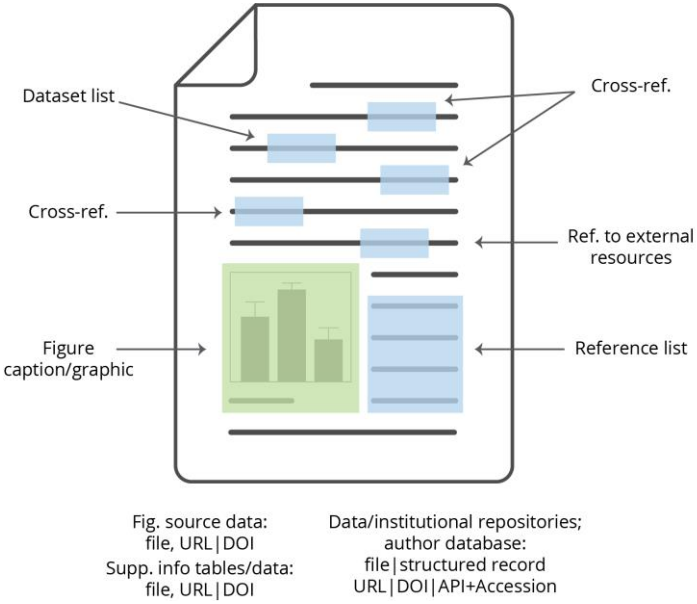
Access more content

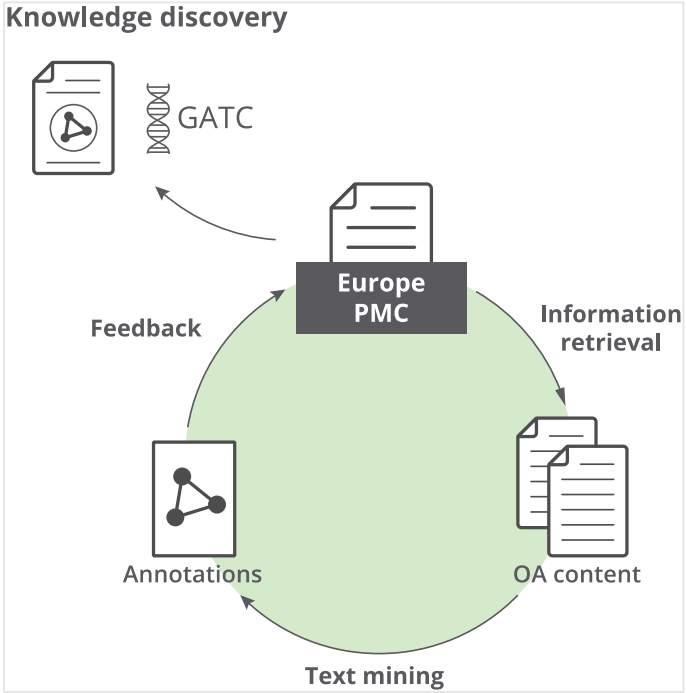


Single search interface



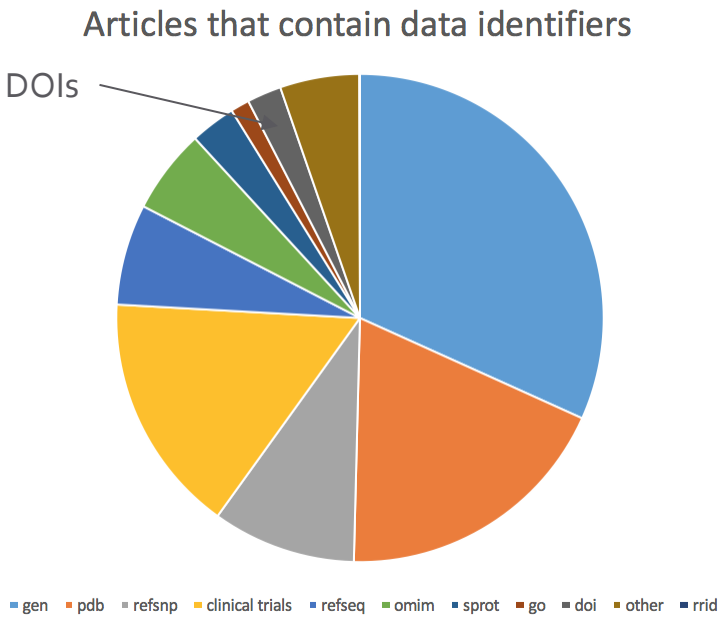
Extracting information from semi-structured data (reading)



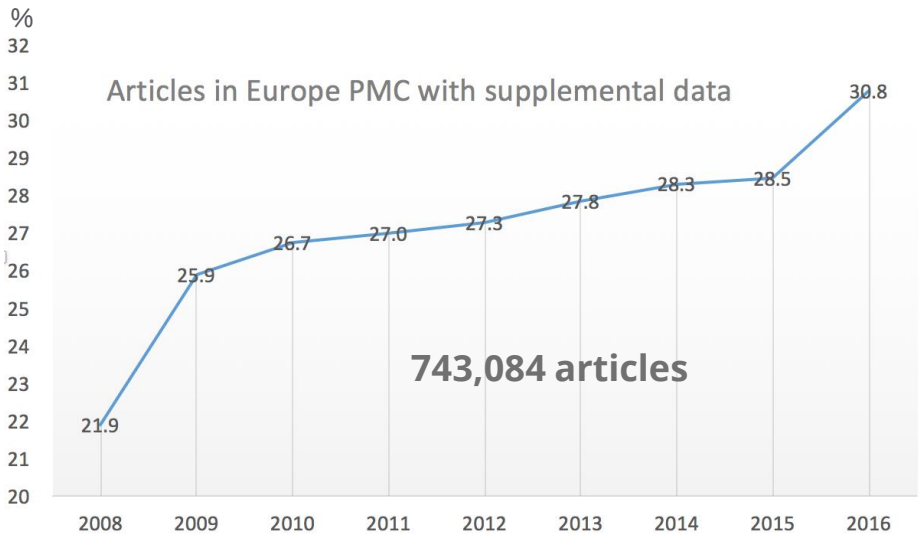


Data "citation" in Europe PMC

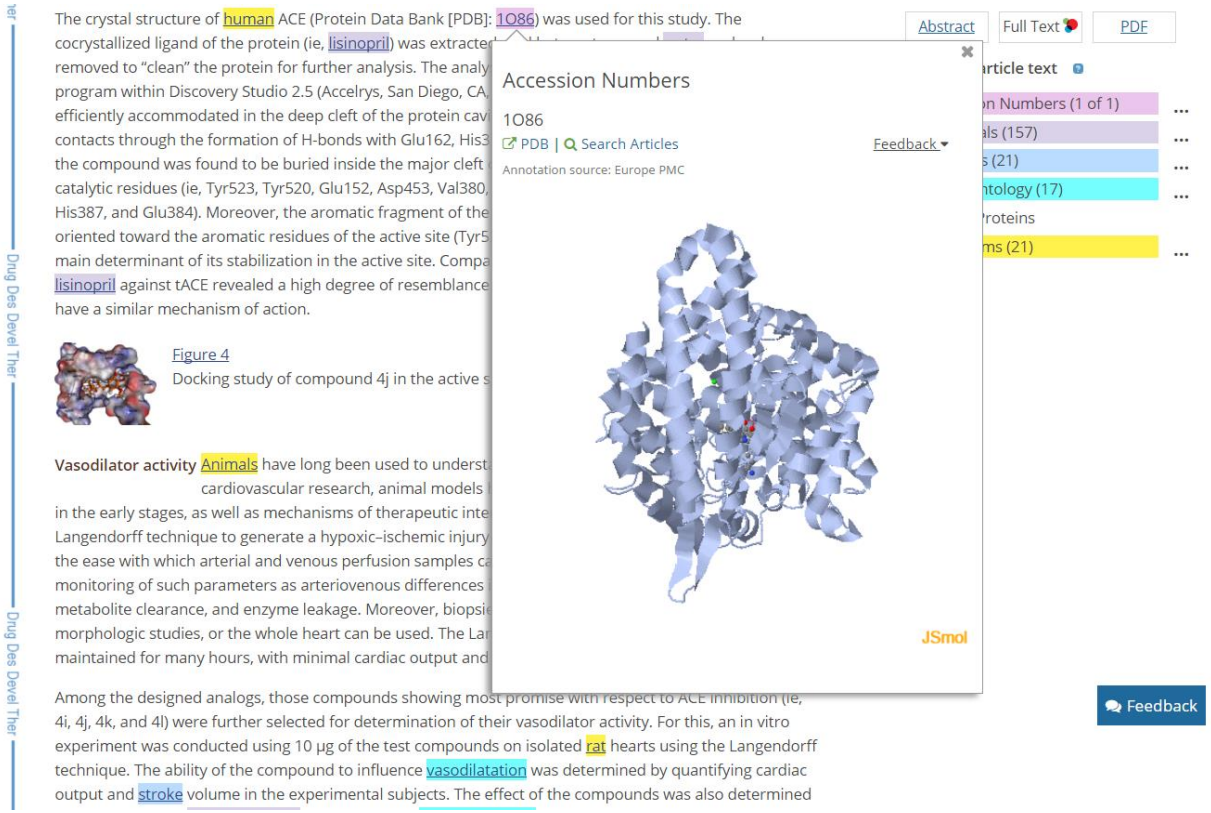
<10% articles mention data identifiers



~30% articles have supplemental data



Annotations on papers



Europe PMC

Authors and ORCIDs

Salt-inducible kinase 3, SIK3, is a new gene associated with hearing.
(PMID:25060954 PMCID:PMC4222365)

Abstract | Citations | BioEntities | Related Articles | External Links

Wolber LE , Girotto G , Buniello A , Vuckovic D , Pirastu N , Lorente-Cánovas B , Rudan I , Hayward C , Polasek O , Ciullo M , Mangino M , Steves C , Concas MP , Cocca M , Spector TD , Gasparini P , Steel KP , Williams FM

Department of Twin Research and Genetic Epidemiolo EH, UK.

Human Molecular Genetics [2014, 23(23):6407-6418]

Type: Journal Article, Meta-Analysis, Research Support
DOI: 10.1093/hmg/ddu346

Frances Williams
Imperial College London
Author Profile
ORCID
Search articles by ORCID
Filter current search by ORCID

Abstract

Hearing function is known to be heritable, but ssoc been identified to date in the adult populatio iatio from the G-EAR consortium and TwinsUK were used for meta-analysis. Hearing ability Northern and Southern European ancestry (n = 4591) and the Silk Road (n = 348) was audiometry and summarized using principal component (PC) analysis. Genome-wide were conducted separately in each sample assuming an additive model adjusted for subjects. Meta-analysis was performed using 2.3 million single-nucleotide polymorph of the three PCs of hearing ability in 4939 individuals. A single SNP lying in intron 6 of (SIK3) gene was found to be associated with hearing PC2 (P = 3.7×10(-8)) and further sequence in a subset. To determine the relevance of this gene in the ear, expression in mouse cochlea of different ages. Sik3 was expressed in murine hair cells during ea the spiral ganglion during early development and adulthood. Our results suggest a d hearing and may be required for the maintenance of adult auditory function.

Formats

Abstract | Full Text | PDF

Show annotations in this article

- Chemicals (1)
- Gene Ontology (8)
- Genes/Proteins (6)
- Organisms (2)

Frances Williams
Imperial College London
Share:

87 Publications in Europe PMC
50 Free full text articles in Europe PMC
2,253 Open citations in Europe PMC

Items published and open citations received each year
Download chart

Abstract only | Free full text | Free full text and licensed for reuse | Open citations of all articles, by year cited | Open citations of full text articles, by year cited

Europe PMC

Claiming Individual Studies

EMBL-EBI

Services | Research | Training | About us

Protein Data Bank in Europe
Bringing Structure to Biology

Search

Examples: hemoglobin, BRCA1_HUMAN

Search EMDB

Feedback

PDBe › 1b0b

X-ray diffraction
1.43Å resolution

HEMOGLOBIN I FROM THE CLAM LUCINA PECTINATA, CYANIDE COMPLEX AT 100 KELVIN

Source organism: *Phacoides pectinatus*

Primary publication:
Cyanide binding to Lucina pectinata hemoglobin I and to sperm whale myoglobin: an x-ray crystallographic study.

Bolognesi M, Rosano C, Losso R, Borassi A, Rizzi M, Wittenberg JB, Boffi A, Ascenzi P

Biophys. J. **77** 1093-9 (1999)
PMID: 10423453

Released: 18 Feb 2000

Model geometry
Fit model/data

Quick links

1b0b overview
- Citations

Structure analysis
Function and Biology
Ligands and Environments
Experiments and Validation

View
Downloads
3D Visualisation

Function and Biology

Details

Biochemical function: heme binding
Biological process: oxygen transport
Cellular component: extracellular region

Sequence domains:
- Erythrocruorin
- Globin
- Globin-like

Structure domain:
- Globins

Ligands and Environments

2 bound ligands:

1 x CYN 1 x HEM

1 modified residue:

1 x SAC

Structure analysis

Details

Assembly composition: monomeric (preferred)
Entry contents: 1 distinct polypeptide molecule
Macromolecule:
Hemoglobin-1

Molecule details ›

Chain: A
Length: 142 amino acids
Theoretical weight: 14.83 KDa
Source organism: *Phacoides pectinatus*
UniProt:
- Canonical: P41260 (Residues: 2-143; Coverage: 99%)

Sequence domains: Globin
Structure domains: Globins

Experiments and Validation

Details

Metric	Percentile Ranks	Value
Rfree		0.181
Clashscore		6
Ramachandran outliers		0
Sidechain outliers		0
RSRZ outliers		0

Worse — Better

Percentile relative to all X-ray structures
Percentile relative to X-ray structures of similar resolution

X-ray source: EMBL/DESY, HAMBURG BEAMLINE BW7A
Spacegroup: $P2_1$

Citations

8 review citations

Redox chemistry and chemical biology of H2S, hydropersulfides, and derived species: implications of their possible biological activity and utility.
Ono et al. (2014)

7 more

3 mentions without citation

Protein secondary structure assignment revisited: a detailed analysis of different assignment methods.
Martin et al. (2005)

2 more

PDB_REDO

The sliders below show the change in model quality between original PDB entry and the PDB_REDO entry

Model Geometry
Fit model/data

PDB-REDO

ORCiD claim

You can sign-in with ORCiD to claim this entry

Remember me on this computer

Europe PMC

Batch Retrospective Claiming

wwwdev.ebi.ac.uk/ebisearch/search.ebi?db=arrayexpress-repository&query=RNA-seq

Apps | E-MTAB-1-heatmap... | Bookmarks | Profile | Add to Mendeley | missing | summary | storage io | Current ISL Errors | RNASeq-er API log | Altmetric it! | dryad

EMBL-EBI | Services | Research | Training | About us | EMBL-EBI Hinxton

EBI Search

RNA-seq

Examples: VAV_HUMAN , tpi1 , Sulston ... Build Query

Help & Documentation | About EBI Search | Feedback

Search results for **RNA-seq**

Showing **15** results out of **10,560** in All results → Gene expression → ArrayExpress

Filter your results

Source

All results (1,990,134)

Gene expression (25,523)

ArrayExpress (10,560)

Repository

ArrayExpress (10,560)

Organisms

Mus musculus (2,737)

Homo sapiens (2,637)

Arabidopsis thaliana (467)

Drosophila melanogaster (426)

Caenorhabditis elegans (291)

Saccharomyces cerevisiae (217)

Save result | Claim to ORCID | Create RSS feed

ArrayExpress (10,560 results)

Genome-wide transcriptome analysis bZIP29 dominant negative repressor line in root tips of Arabidopsis

... responses. **RNA-seq** transcriptome profiling of the root meristem shows that bZIP29 target genes are linked ...

Related data

Source: ArrayExpress
ID: E-MTAB-3755

Single-cell **RNA-seq** analysis of human pancreas from healthy individuals and type 2 diabetes patients

We used single-cell **RNA**-sequencing to generate transcriptional profiles of endocrine and exocrine cell types of the human pancreas. Pancreatic tissue and islets were obtained from six healthy and four T2D cadaveric donors. Islets were cultured and dissociated into single-cell suspension. Viable ...

Related data

Source: ArrayExpress
ID: E-MTAB-5061

Europe PMC

All EMBL-EBI Data Claims to ORCID

EBI Search

domain_source:orcid_data_claims

Examples: VAV_HUMAN , tpi1 , Sulston ...

Build Query

Help & Documentation | About EBI Search | Feedback

Search results for **domain_source:orcid_data_claims**

Showing **15** results out of **15** in All results → Samples & ontologies → ORCID data claims

Filter your results

Source

All results (15)

Samples & ontologies (15)

ORCID data claims (15)

Dataset type

Metabolights (15)

Save result

Create RSS feed

ORCID data claims (15 results)

MTBLS372

Metabolomic profiling of Fraxinus excelsior genotypes tolerant or susceptible to ash dieback disease reveals changes in iridoid glycosides

ORCID(s): 0000-0002-7219-0398

Related data

Source: ORCID data claims
ID: MTBLS372

MTBLS415

3D DESI mass spectrometry imaging of 52 serial sections from a human colorectal adenocarcinoma

ORCID(s): 0000-0002-1007-317X

Related data

Source: ORCID data claims
ID: MTBLS415

MTBLS200

Temporal characterization of serum metabolite signatures in lung cancer patients undergoing treatment

ORCID(s): 0000-0003-3674-7336

Related data

Source: ORCID data claims
ID: MTBLS200

Synergistic Efforts at EMBL-EBI
http://www.omicsdi.org/search

Home | Browse | API | Database | Help | Login | *:* | Advanced | Search

Yasset Perez-Riverol

I'm a Project Leader of Multiomics at the EMBL-European Bioinformatics Institute (Hinxton, Cambridge, UK). I earned undergraduate degrees in Software Engineer (2006) and a doctoral degree in Biochemistry (2013) from the University of Havana. After finishing my PhD in Havana he joined the PRIDE team in 2014. I have lead several development projects such as PRIDE Inspector Toolsuite, and Omics Discovery Index a major resource to find, discovery and link omics datasets.

Contact Info

EMBL-EBI

yperez@ebi.ac.uk

https://orcid.org/0000-0001-6579-6941

0000-0001-6579-6941

Yasset Perez-Riverol

Datasets

PIA - Mouse Benchmark Dataset (150 | 0 | 0 | 0)

This Dataset is no actual new study but the mouse benchmark dataset used in the PIA manuscript.
ORGANISM(S): Mus musculus

2015-05-08 | PXD000790 | Pride

mouse | benchmark | Reference | Biomedical | Cite

PIA - Yeast Gold Standard Benchmark Dataset (159 | 0 | 0 | 0)

This Dataset is no actual new study but the Yeast Gold Standard benchmark dataset used in the PIA manuscript.
ORGANISM(S): Saccharomyces cerevisiae

2015-05-08 | PXD000792 | Pride

yeast | benchmark | Reference | Cite

PIA - iPRG2008 Benchmark Dataset (198 | 0 | 0 | 0)

This dataset is no actual new study but the iPRG2008 benchmark dataset used in the PIA manuscript.
ORGANISM(S): Mus musculus

2015-05-08 | PXD000793 | Pride

mouse | iPRG2008 | benchmark | Technical | Reference | Cite

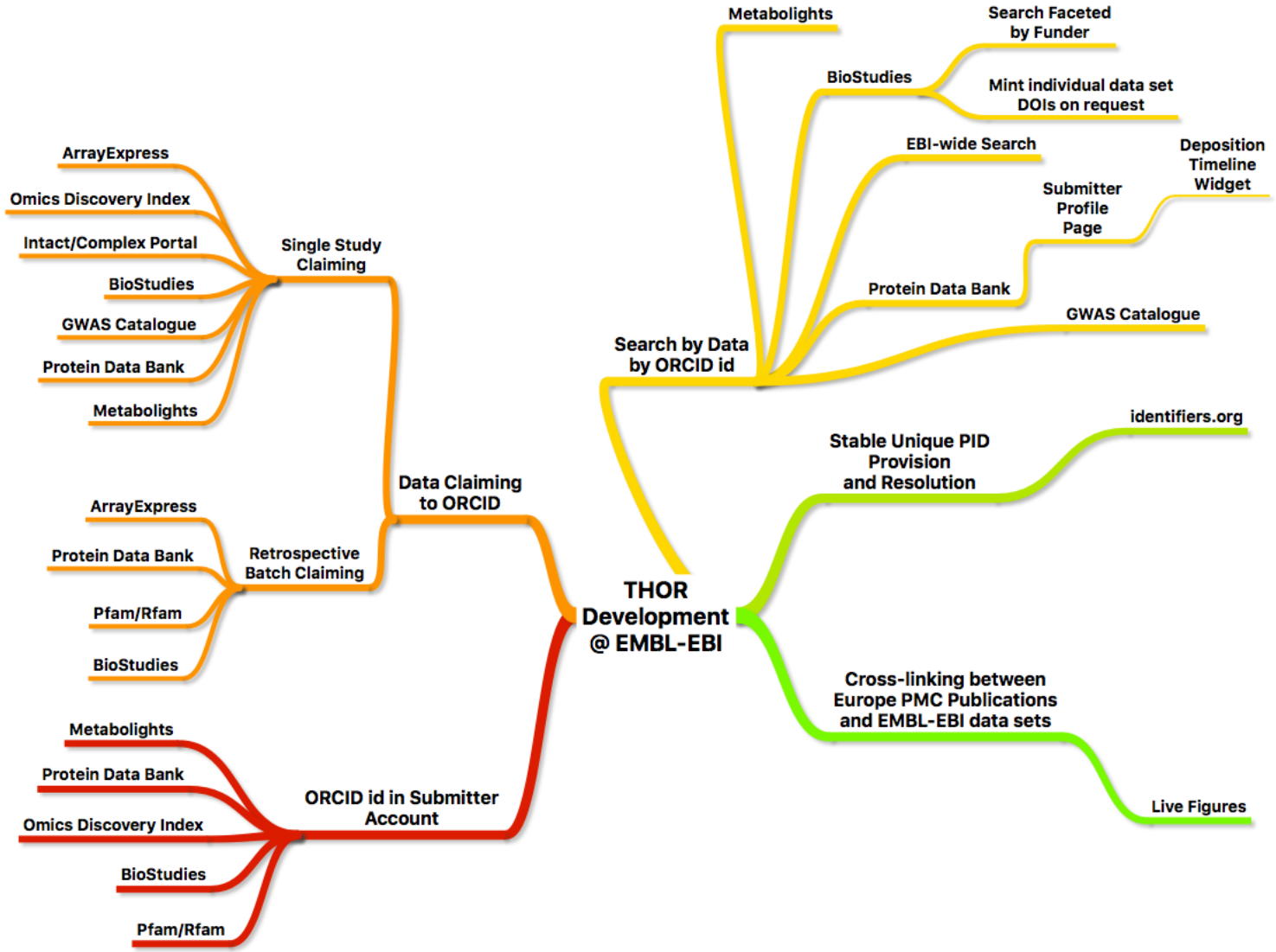
Search Europe PMC
By ORCID

EBI ORCID HUB

ORCID

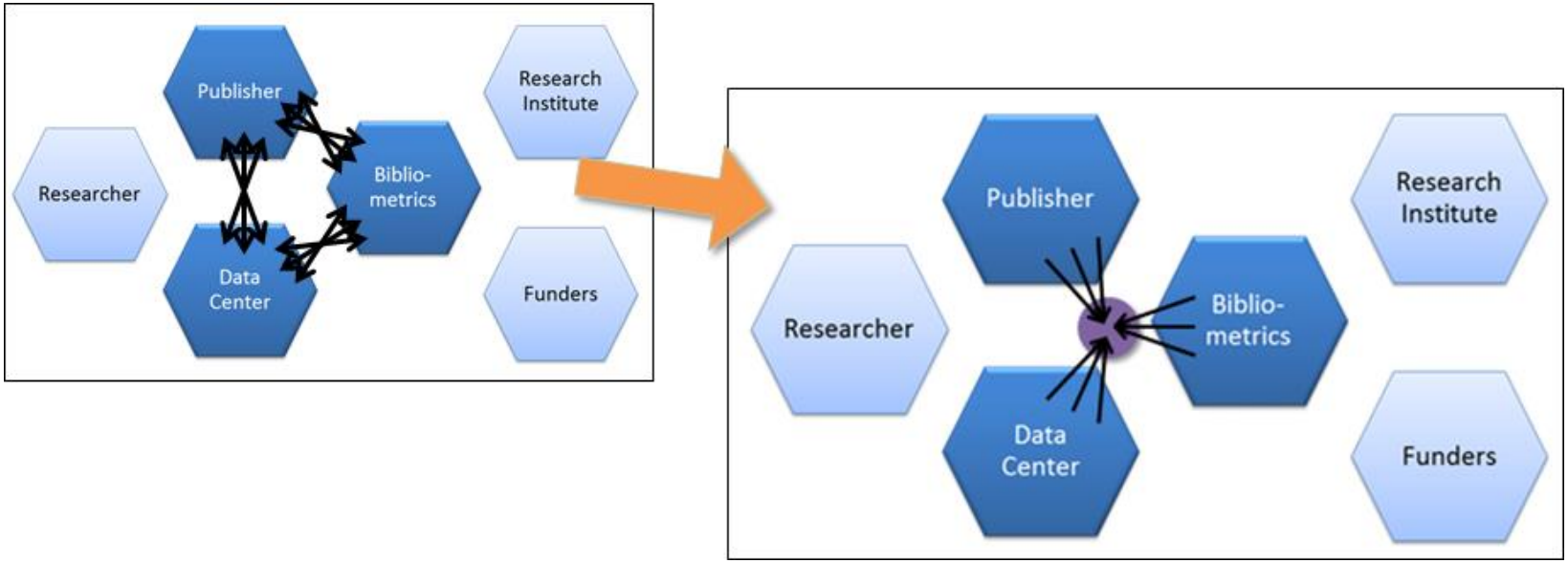
Europe PMC

EMBL-EBI THOR Development – At a Glance



Scholix: Data-Literature Links exchange

- Harmonized format for Data-Literature links
- Between natural link hubs
  - CrossRef
  - DataCite
  - OpenAIRE



Scholix at Europe PMC

- Different origin Data-Literature links in various places in API and User Interface
- Consolidation into one API method providing links in Scholix format

Connecting with healthcare providers at diagnosis: adolescent/young adult cancer survivors' perspectives.
(PMID:28617094 PMCID:PMC5510205)

Abstract | Citations | BioEntities | Related Articles | External Links

Phillips CR[1], Haase JE[1], Broome ME[2], Carpenter JS[1], Frankel RM[3]

Affiliations

International Journal of Qualitative Studies on Health and Well-being [01 Dec 2017, 12(1):1325699]

Type: research-article, Journal Article
DOI: 10.1080/17482631.2017.1325699

Abstract

Adolescents and young adults (AYAs) with cancer are a vulnerable and underserved population. AYAs' cancer survivorship is complicated by physical and psychosocial late effects which requires long-term follow-up. Connectedness with healthcare providers (HCPs) is a protective factor that may improve long-term follow-up behaviours of AYAs. However, little is known about AYAs' experiences connecting with HCPs. The purpose of this study was to describe AYA cancer survivors' experiences connecting with HCPs. This empirical phenomenological study interviewed nine AYA cancer survivors diagnosed during adolescence. Individual interviews were conducted and analysed using an adapted Colaizzi approach. The essential structure reveals that AYAs begin their experience of connectedness with a sense of disconnectedness prior to treatment. The diagnosis is a period of confusion and emotional turmoil that interfere with the AYAs' ability to connect. When AYAs come to accept their illness and gain familiarity with the environment, they then put forth an effort to

Data in different Places

- Connecting with healthcare providers at diagnosis: adolescent/young adult cancer survivors' perspectives.
  (PMID:28617094 PMCID:PMC5510205)

Abstract | Citations | BioEntities | Related Articles | External Links

Gene Ontology (GO) Terms

Identified 3 unique GO Terms in the full text

- behaviours (16)
- cell (1)
- sleep (1)

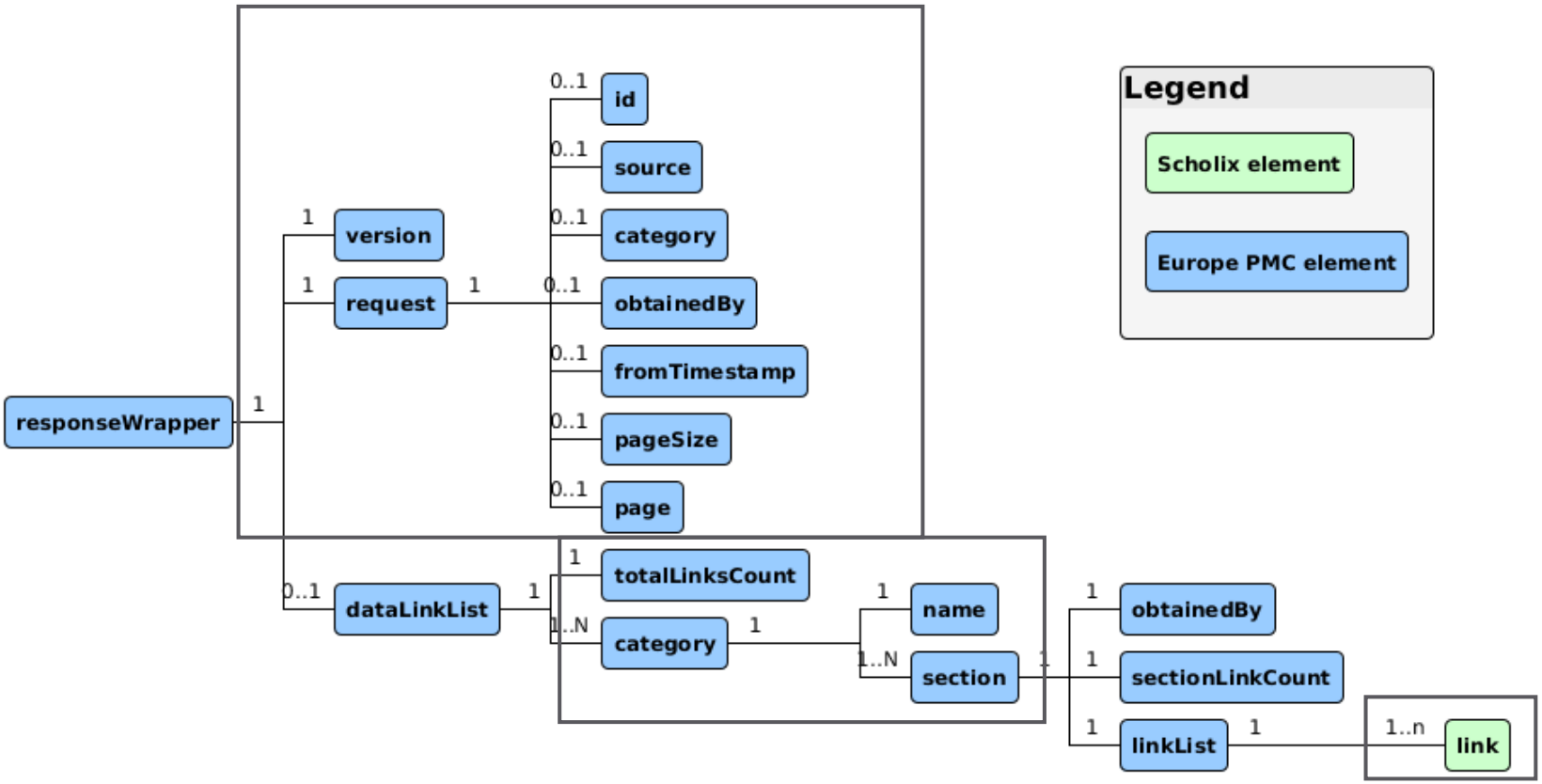
Show all items

Species

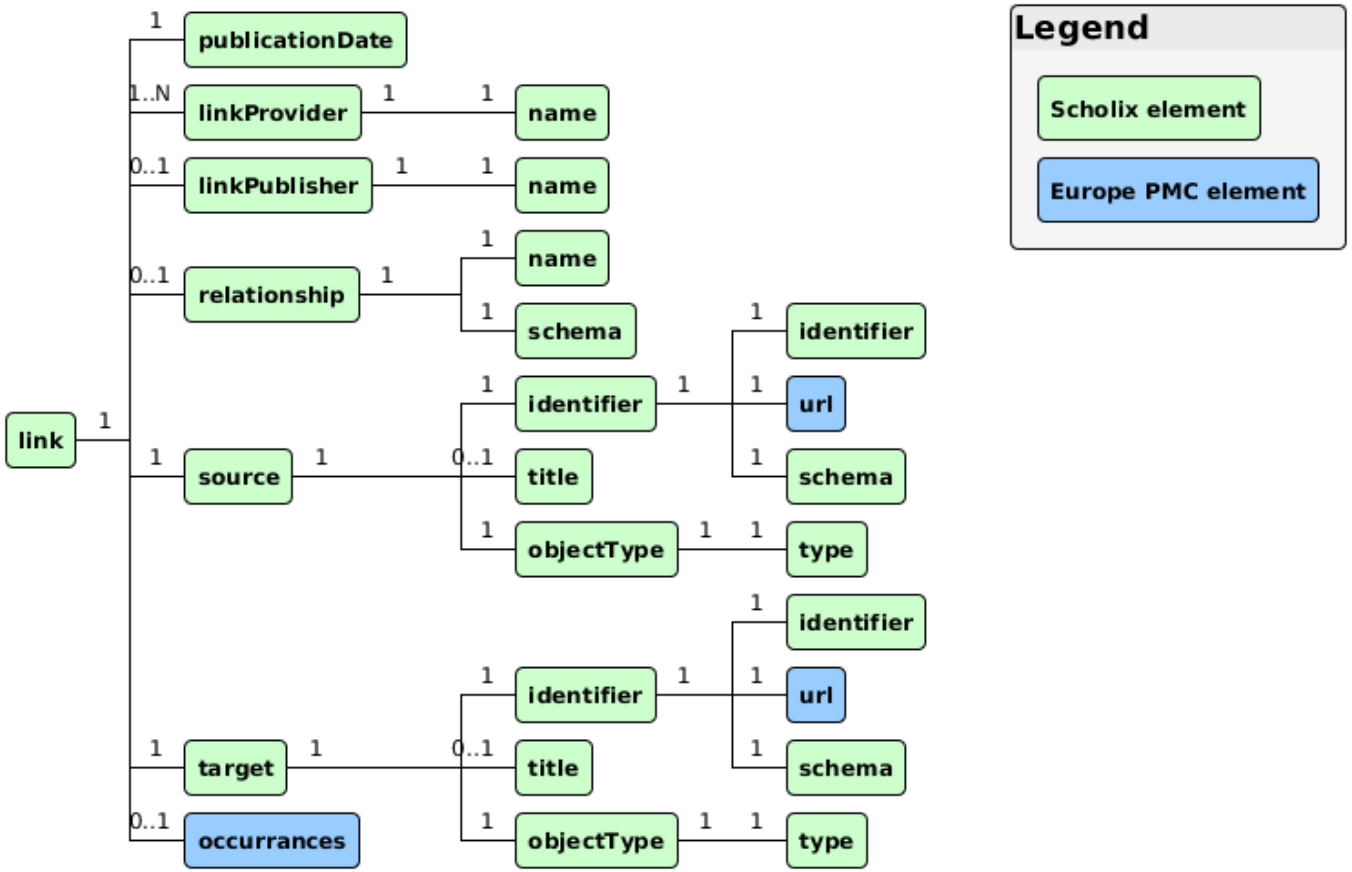
Identified 1 unique Species in the full text

- morphine (1)

Response Structure



Scholix link format+



Scholix at Europe PMC

Europe PMC | About | Tools | Developers | Help | Europe PMC plus

Search worldwide, life-sciences literature

Search | Advanced Search

E.g. "breast cancer" HER2 Smith J

A tick salivary protein targets cathepsin G and chymase and inhibits host inflammation and platelet aggregation.
(PMID:20940421 PMCID:PMC3031492)

Abstract | Citations | Related Articles | Data | BioEntities | External Links

Data behind this article

BioStudies. Primary data and supplemental files
http://www.ebi.ac.uk/biostudies/studies/S-EPMC3031492

Figures are available in the full text of the article

Data associated with this article

4 UniProt records that cite this article
1 PDBe record that cites this article
4 ENA records that cite this article
2 OMIM records that cite this article

Recent Activity | Export | Tweet

Formats

Abstract | Full Text

Cited by 54 | view all

Supported by:

Supported by:

More questions? Contact us

https://europepmc.org/
@EuropePMC_News

Questions about Europe PMC
Helpdesk@EuropePMC.org

My contact details
graf@ebi.ac.uk