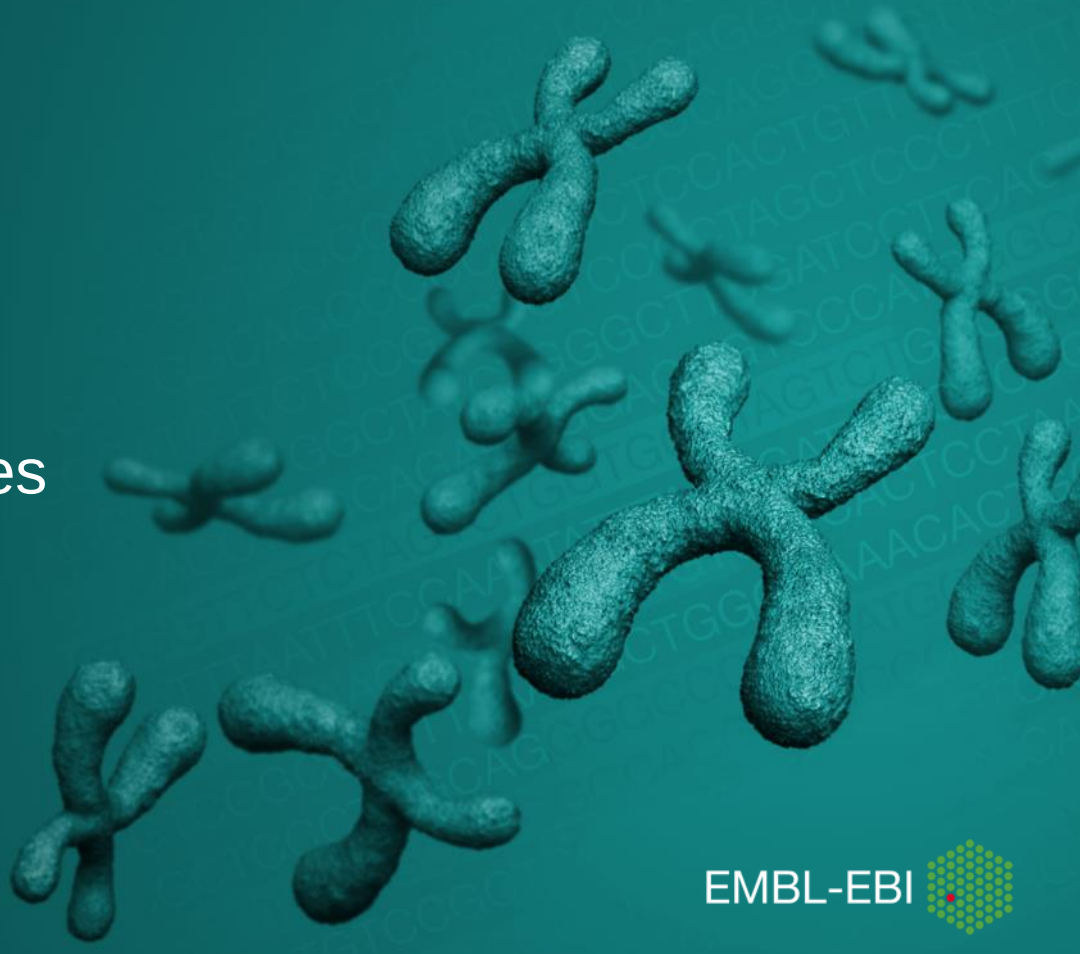
# A tour of the Europe PMC full text database and related text-based tools at the EMBL-EBI

Jo McEntyre

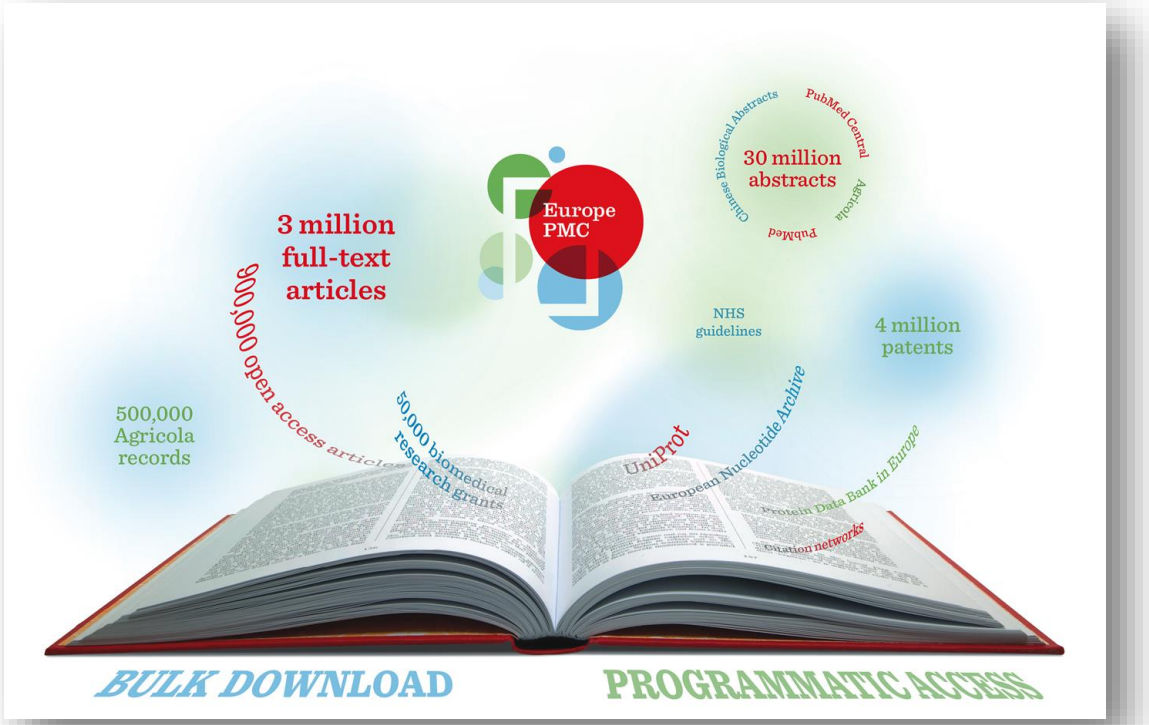
Head of Literature Services

www.ebi.ac.uk



EMBL-EBI

# Europe PMC Overview




- 30 million documents of which 3 million full text
- Enrichments: ORCIDs, citations, named entities, DOIs, data links
- Website, web services, FTP
- 29 funders including EMBL
- Over 1M IPs per month

- A PMCI partner

Europe PubMed Central

"mammal diversity"

## Results

Recent Activity | Export

Results (130) Sort by: Relevance | Date | Times Cited

1 2 3 4 5 6

Results 1 - 25 of 130

Benefits to poorly studied taxa of conservation of bird and **mammal diversity** on islands.
(PMID:25065901)
Aslan C, Holmes N, Tershy B, Spatz D, Croll DA
Conserv Biol [2015, 29(1):133-142]
Cited: 0 times

Higher speciation and lower extinction rates influence **mammal diversity** gradients in Asia.
(PMID:25648944 PMCID:PMC4333168) Free full text article
Tamma K, Ramakrishnan U
BMC Evol Biol [2015, 15:11]
Cited: 0 times

Reading **Mammal Diversity** from Flies: The Persistence Period of Amplifiable Mammal mtDNA in Blowfly Guts (Chrysomya megacephala) and a New DNA Mini-Barcode Target.
(PMID:25898278 PMCID:PMC4405593) Free full text article

### Popular content sets

Full Text articles only (85)

Open Access articles only (53)

All reviews (6)

# What do Pneumocystis organisms tell us about the phy their hosts? The case of the woodmouse Apodemus sy continental Europe and western Mediterranean islands

(PMID:25830289 PMCID:PMC4382281)

Abstract | Citations | BioEntities | Related Articles | E

Demanche C, Deville M, Michaux J, Barriel V, Pinçon C, Aliouat-Denis CM, Pottier M Aliouat el M, Dei-Cas E, Morand S, Guillot J

Laboratoire de Parasitologie (EA4547), Faculté de Pharmacie, Université de Lille, Li de Lille, Centre d'Infection et d'Immunité de Lille, Inserm U1019, UMR CNRS 8204, d'Alfort, Biologie et Diversité des Pathogènes Eucaryotes Emergents, Lille, France.

PloS one [2015, 10(4):e0120839]

**Type:** Journal Article, Research Support, Non-U.S. Gov't
**DOI:** 10.1371/journal.pone.0120839

**Abstract**

Gene Ontology(1) Genes

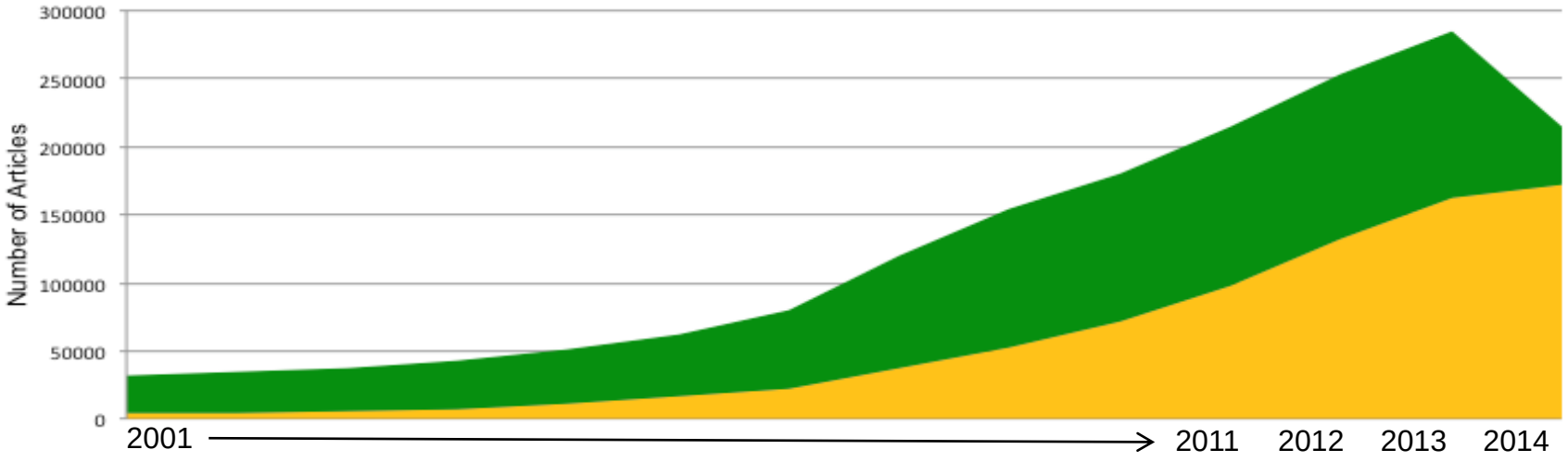
Pneumocystis fungi represent a highly diversified biological group with nume display a strong host-specificity suggesting a long co-speciation process. In presence and genetic diversity of Pneumocystis organisms was investigated from woodmice (Apodemus sylvaticus) collected on western continental Euro islands. The presence of Pneumocystis DNA was assessed by nested PCR a

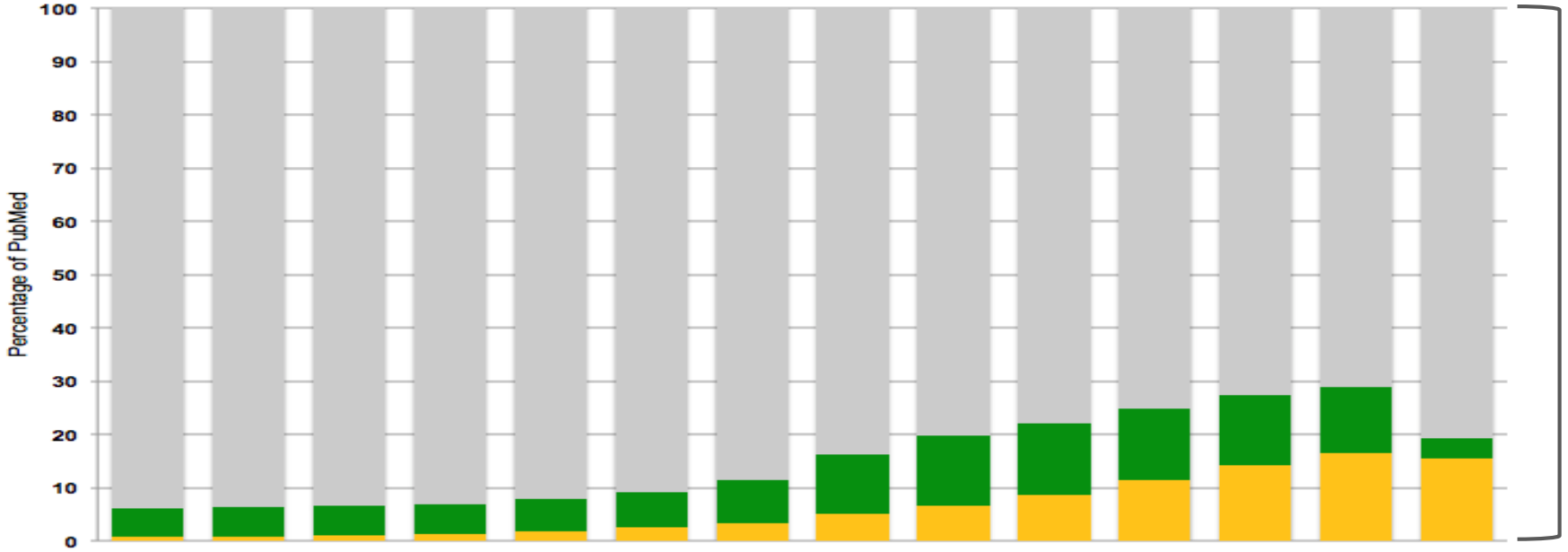
## Species

Identified 20 unique Species in the full text

- Pneumocystis (152)
- Apodemus sylvaticus (33) (A. sylvaticus)
- Rodentia (10) (rodents)
- mice (9) (mouse)
- Heligmosomoides polygyrus (8) (H. polygyrus)
- nematode (7)
- Animals (7)
- mammals (6)
- fungi (4)
- primate (4)
- Apodemus (4)
- P. murina (4)
- Muridae (2)
- Trichuris muris (1)
- Turkey (1)
- humans (1)
- Tadarida brasiliensis (1)
- S. aestuans (1)
- macaques (1)
- rats (1)

# Europe PMC Content



# Unique Identifiers

ORCID

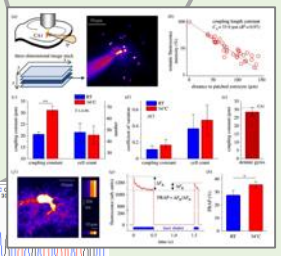

PMID, PMCID, DOI

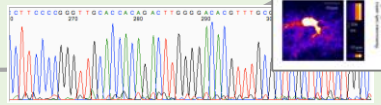
Ringgold ID

Reused from: seier+seier, Flickr

Reused from: Images Money, Flickr

GrantID





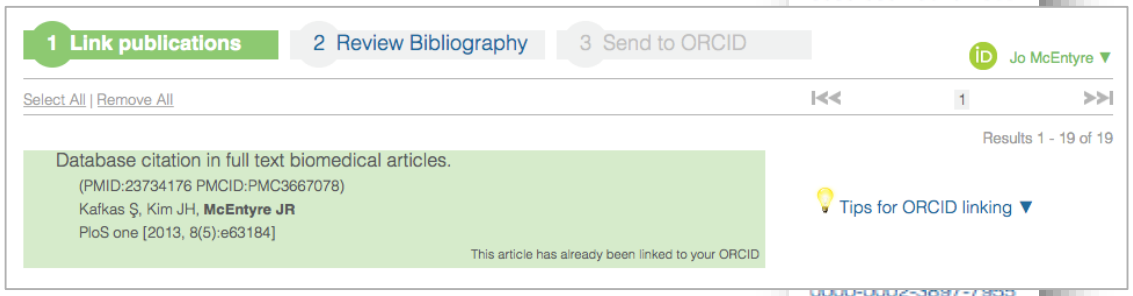
Accession number
DOI, handle

Critical for non-ambiguous integration, impact assessment and credit systems

# (1) Europe PMC and ORCID

## Use of the claiming tool

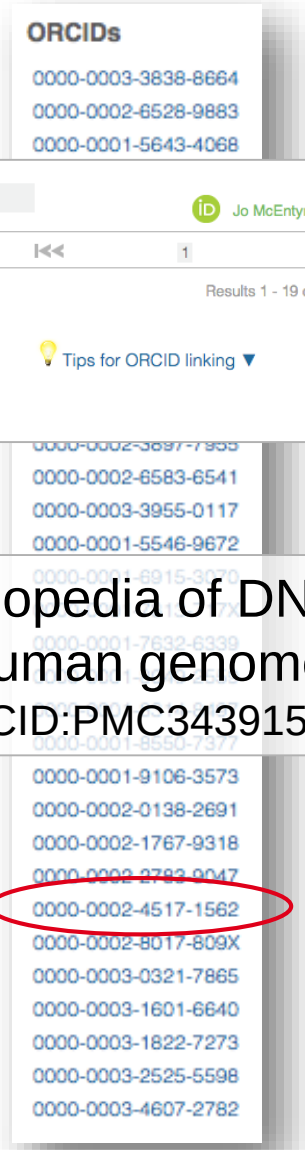
- ~500K article claims
- ~ 25K unique ORCIDs




## Incorporation into Europe PMC

- 1.6M unique articles in Europe PMC linked to ORCIDs
- 115,000 ORCIDs represented

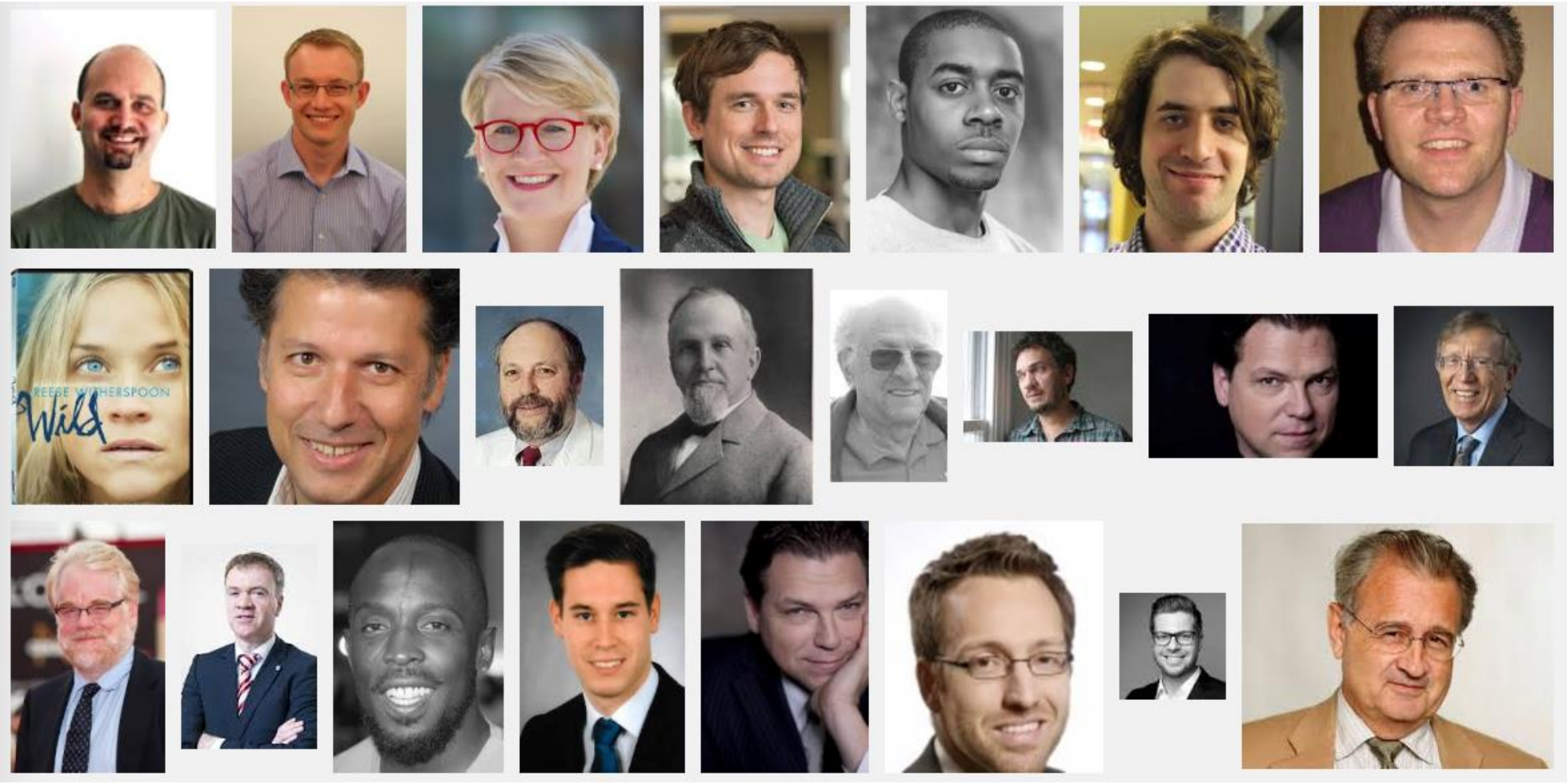
An integrated encyclopedia of DNA elements in the human genome.
(PMID:22955616 PMCID:PMC3439153)




MM Hoffman
15 articles
(74 by name search)



EMBL-EBI


REESE WITHERSPOON
Wild

Found 10 UniProt record(s) citing this article

CG3595

(UniProt:F6

Cytochro

(UniProt:P1

Cytochro

(UniProt:Q9

Cytochro

(UniProt:P0

Myosin re

(UniProt:P4

NADH-ub

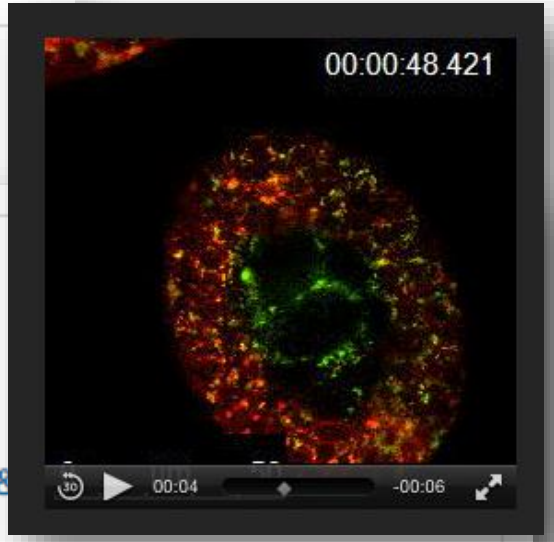
(UniProt:Q9

NADH-ub

## FlyBase

A Database of Drosophila Genes & Genomes

- http://flybase.org/cgi-bin/uniq.html?db=fbrf&field=pubmed_id&




## BioStudies

Database of biological studies

- Female and male gamete mitochondria are distinct and comple
  structure, and genome function
  http://www.ebi.ac.uk/biostudies/studies/S-EPMC3814205

Identified

pdb 3M9

pdb 1ZO

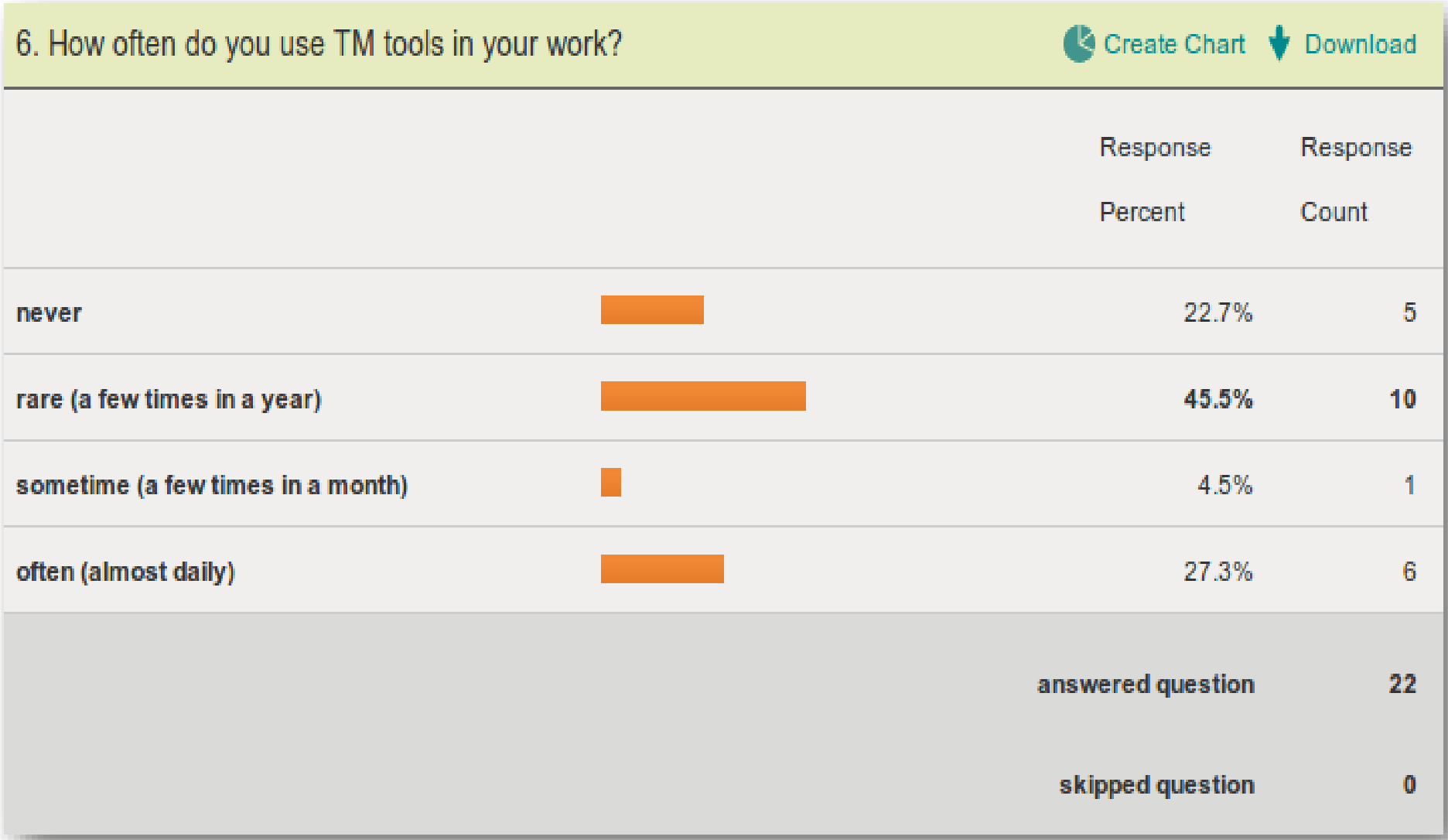
pdb 1QC

pdb 1V54 (1) View Structure


Wikipedia links : 300,000 + articles

EMBL-EBI

# A workshop for curators and text miners: Hinxton, July 2013



6. How often do you use TM tools in your work?

Create Chart Download

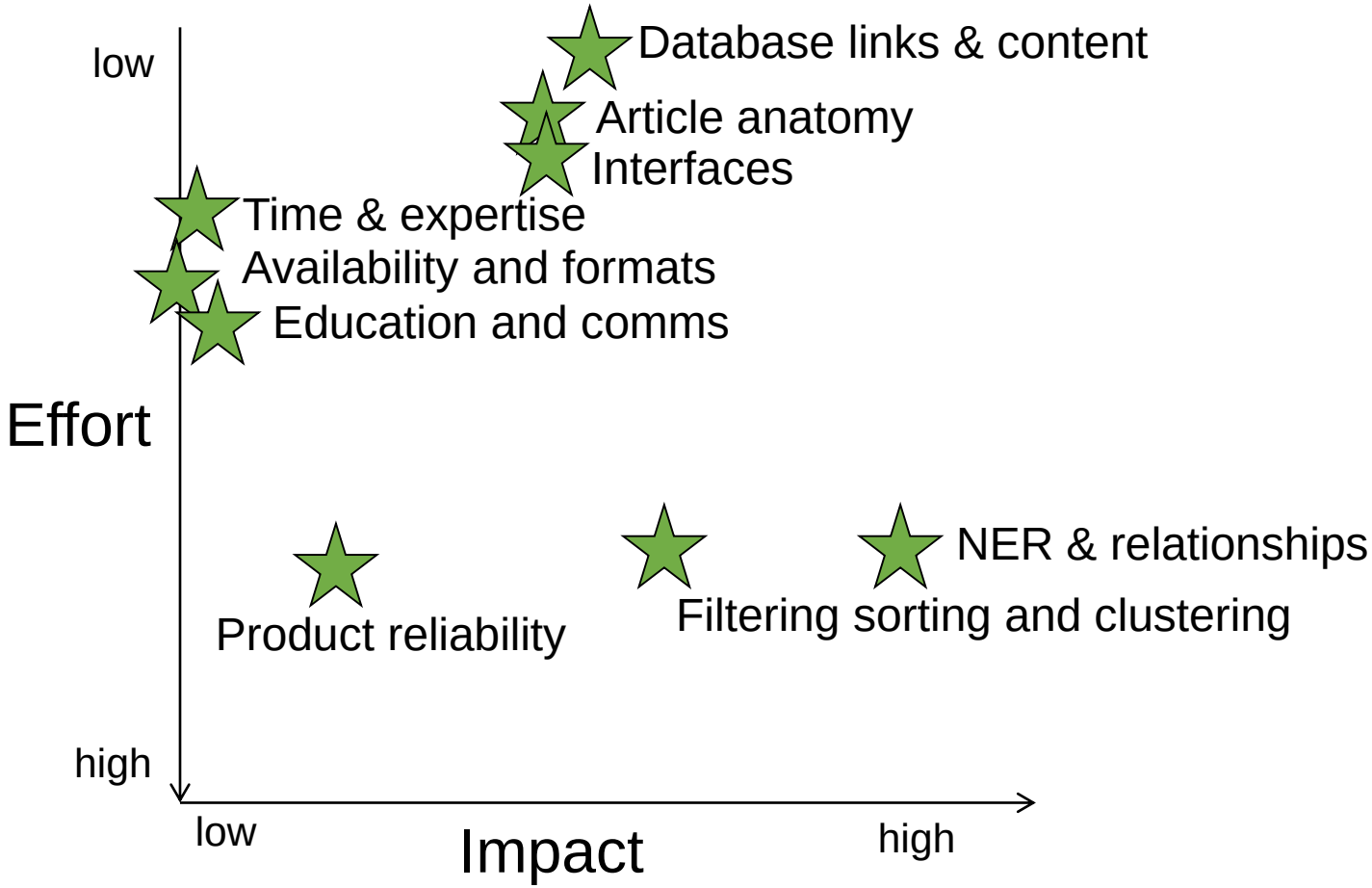
| | Response Percent | Response Count |
|---|---|---|
| never | 22.7% | 5 |
| rare (a few times in a year) | 45.5% | 10 |
| sometime (a few times in a month) | 4.5% | 1 |
| often (almost daily) | 27.3% | 6 |
| answered question | | 22 |
| skipped question | | 0 |

3. How important is text mining currently to services provided by the EBI/Sanger?

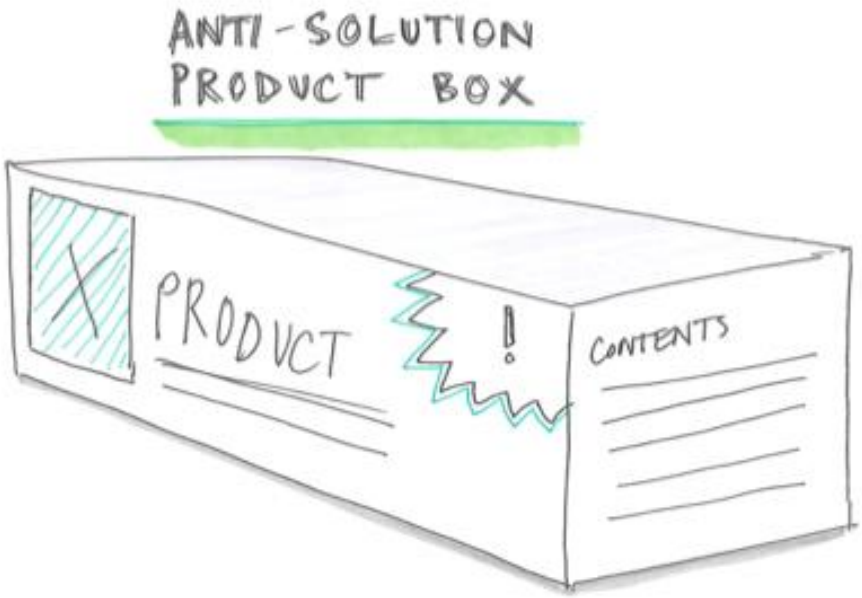
Create Chart Download

| | Response Percent | Response Count |
|---|---|---|
| **not important** | 15.0% | 3 |
| **moderate** | **60.0%** | **12** |
| **important** | 15.0% | 3 |
| **very important** | 10.0% | 2 |
| **answered question** | | **20** |
| **skipped question** | | **2** |

# When asked about text mining features, effort and impact



# When asked about an anti-product




- Reliability
- Performance and speed
- Documentation, guarantees,
- Not being goofy
- Trustworthiness: doing what it says on the tin; not “I feel lucky”
- Easy to use - nice interfaces
- High precision
- Broad appeal
- Not too many results

# Text mining

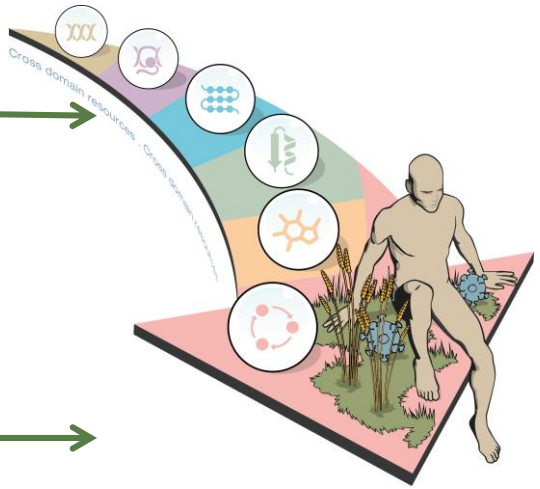
## Literature


## Ontologies

GO

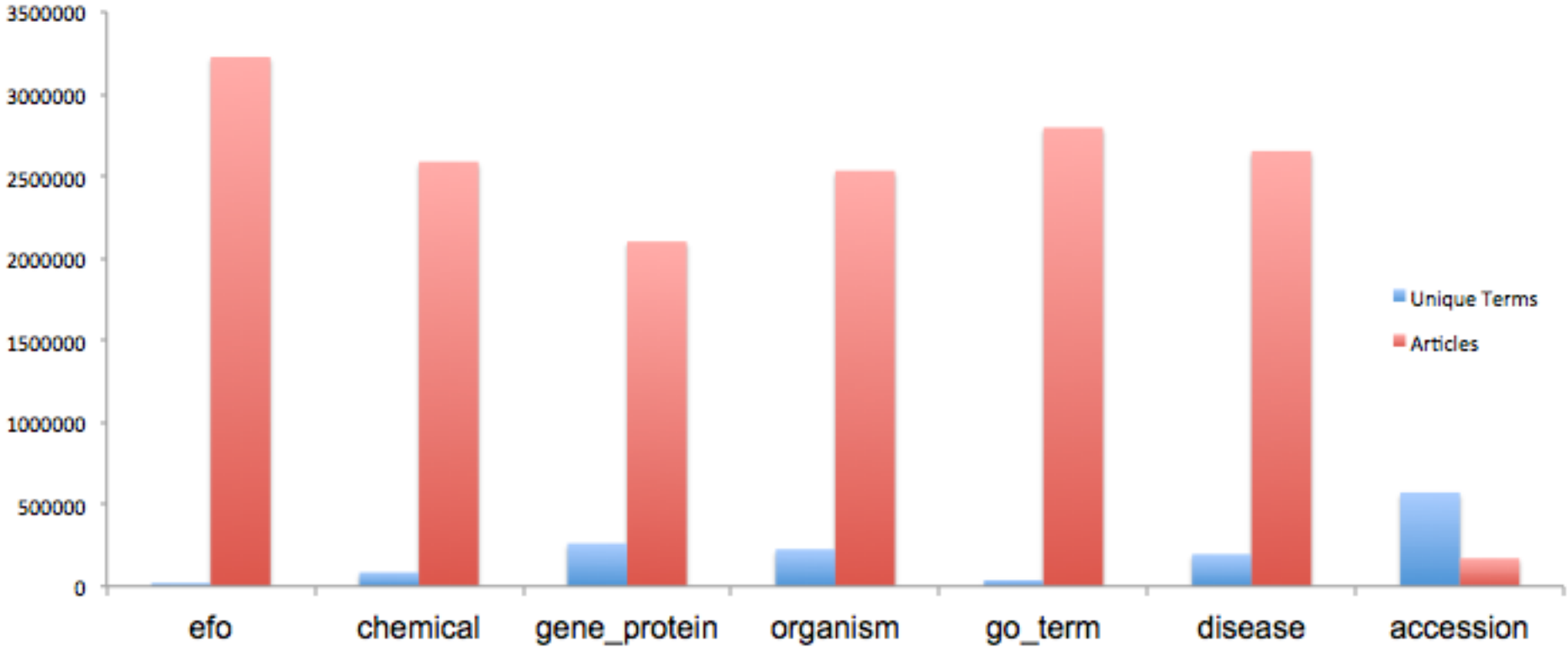
EFO

## Data



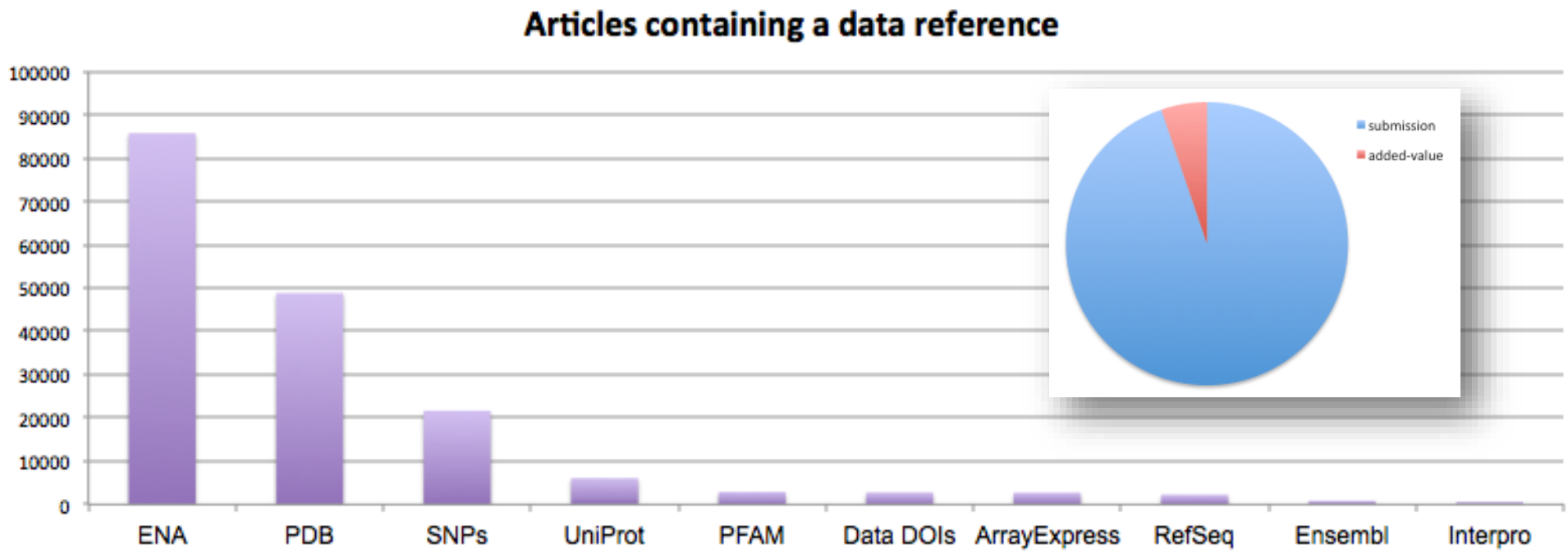

EMBL-EBI

# Scope of routine text mining: full text




Almost a billion unique annotations
Vocabulary mapping and management

# Text-mining data citations: Impact for ELIXIR




Distribution of database citations in the supplementary data of the top 5% articles by databases

| Database | Total number of articles containing database citations in their supplementary data | % of database citations in the supplementary data of the top 5% articles |
|---|---|---|
| ENA | 2,458 | 88.78% |
| PDBe | 1,274 | 86.36% |
| RefSNP | 1,167 | 95.05% |
| UniProt | 1,059 | 83.87% |
| RefSeq | 721 | 63.39% |
| Pfam | 617 | 70.15% |
| InterPro | 499 | 72.46% |
| Ensembl | 377 | 67.62% |
| ArrayExpress | 66 | 88.35% |
| OMIM | 57 | 63.79% |

A reasonable number of articles cite data in supplemental data files, especially as large datasets

# PDBe › 1cbs

## CRYSTAL STRUCTURE OF CELLULAR RETINOIC-ACID-BINDING PROTEINS I AND II IN COMPLEX WITH ALL-TRANS-RETINOIC ACID AND A SYNTHETIC RETINOID

**Source organism:** *Homo sapiens* [9606]

**Primary publication:**

Crystal structures of cellular retinoic acid binding proteins I and II in complex with all-trans-retinoic acid and a synthetic retinoid.

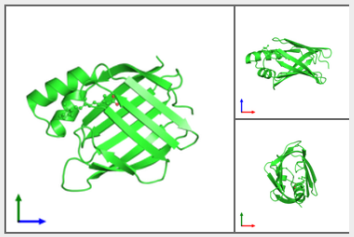
Kleywegt GJ, Bergfors T, Senn H, Le Motte P, Gsell B, Shudo K, Jones TA

*Structure* **2** 1241-58 (1994)
PMID: 7704533

**X-ray diffraction**
**1.8Å resolution**

**Released:** 26 Jan 1995

Model geometry
Fit model/data



Quick links
1cbs overview
Citations
Structure analysis
Function and Biology
Ligands and Environments
Experiments and Validation
View
Downloads
3D Visualisation

## Function and Biology

Details

**Biochemical function:** retinoid binding
**Biological process:** transport
**Cellular component:** extracellular vesicular exosome

**Sequence domains:**
- Cytosolic fatty-acid binding [IPR000463]
- Calycin-like [IPR011038]
- Lipocalin/cytosolic fatty-acid binding domain [IPR000566]
- Calycin [IPR012674]

**Structure domain:**
- Fatty acid binding protein-like

## Structure analysis

Details

**Entry contents:** 1 distinct polypeptide molecule
**Assemblies:** homomeric monomer
**Polymer:**
Cellular retinoic acid-binding protein 2

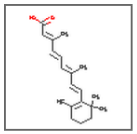
Molecule details ›

**Chain:** A
**Length:** 137 amino acids
**Theoretical weight:** 15.58 KDa
**Source organism:** *Homo sapiens* [9606]
**Expression system:** *Escherichia coli BL21(DE3)* [469008]
**UniProt:** P29373
**Gene name:** CRABP2
**Sequence domains:** Lipocalin / cytosolic fatty-acid binding

## Ligands and Environments

Details

**1 bound ligand:**

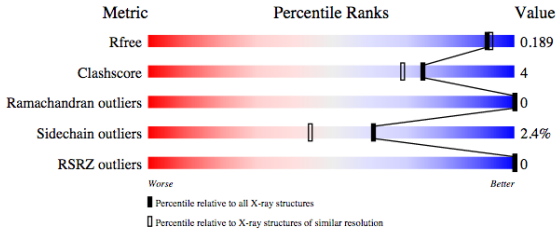


1 x REA

**No modified residues**

## Experiments and Validation

Details




**Spacegroup:** $P2_12_12_1$

**Unit cell:** a: 45.65Å b: 47.56Å c: 77.61Å α: 90° β: 90° γ: 90°

**R-values:** R R work R free

## Citations

### 9 review citations

The photochemical determinants of color vision: revealing how opsins tune their chromophore's absorption wavelength.
Wang et al. (2014)

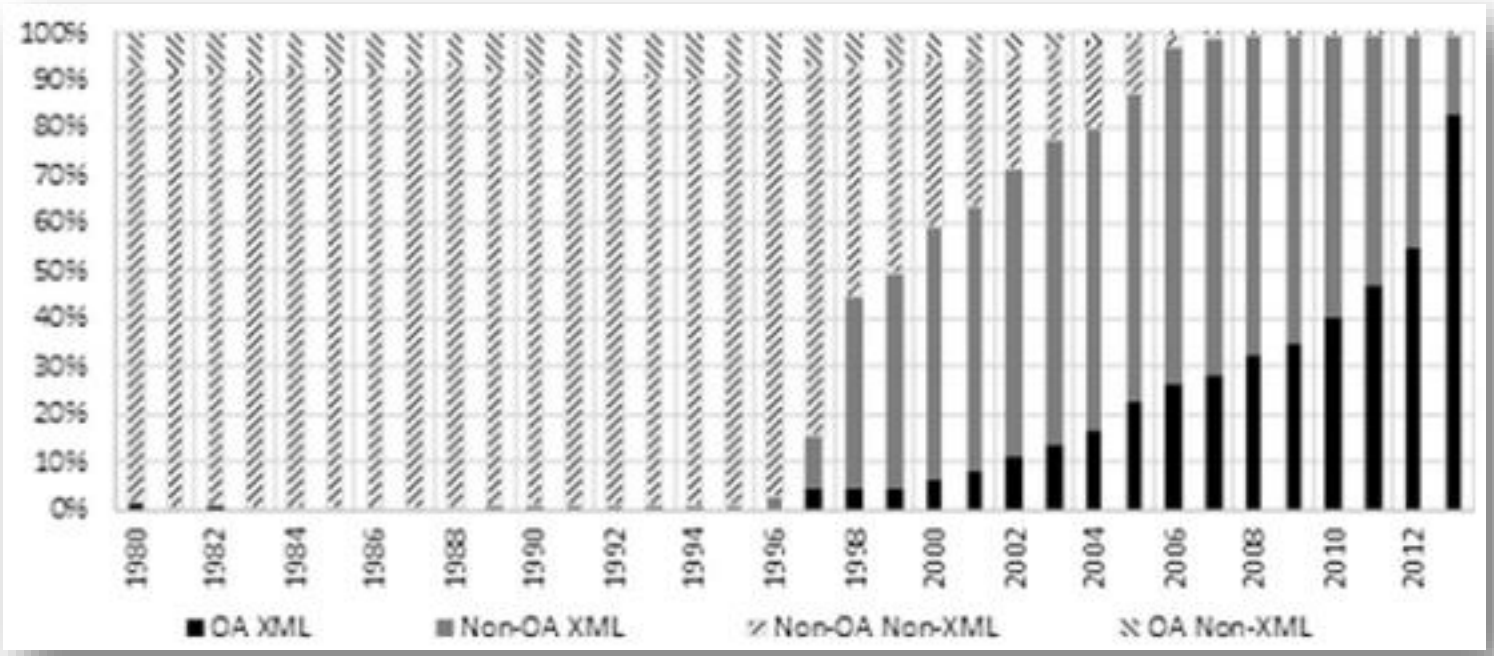
8 more

### 5 mentions without citation

PDBe: Protein Data Bank in Europe.
Velankar et al. (2012)

4 more

# Section Tagger is rule-based and works on XML



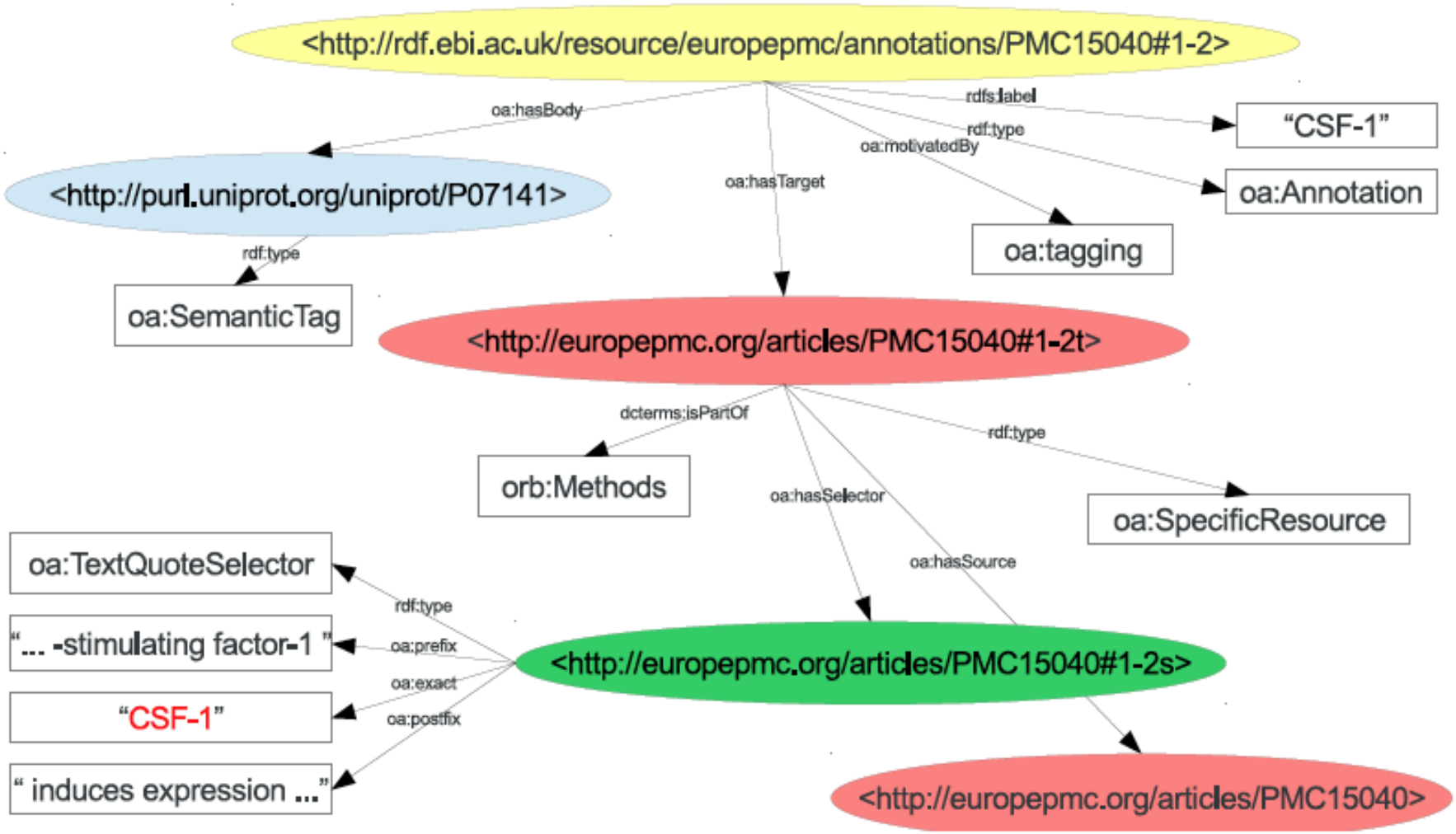

*Conclusion & Future Work*: (conclusion | key message | future | summary | recommendation | implications for clinical practice | concluding remark)".

Precision: 99.84%, Recall: 96.27%, F-score: 98.02%

# Beyond the BioEntities Tab

- More connectivity
- More context for links
- Sharable annotations

- Challenges of scale: nearly a billion annotations generated
- Using Web Annotation Data model


<http://rdf.ebi.ac.uk/resource/europepmc/annotations/PMC15040#1-2>
oa:hasBody
rdfs:label
"CSF-1"
rdf:type
oa:motivatedBy
oa:Annotation
<http://purl.uniprot.org/uniprot/P07141>
oa:hasTarget
oa:tagging
rdf:type
oa:SemanticTag
<http://europepmc.org/articles/PMC15040#1-2t>
dcterms:isPartOf
rdf:type
orb:Methods
oa:hasSelector
oa:SpecificResource
oa:hasSource
oa:TextQuoteSelector
rdf:type
"... -stimulating factor-1 "
oa:prefix
<http://europepmc.org/articles/PMC15040#1-2s>
oa:exact
"CSF-1"
oa:postfix
" induces expression ..."
<http://europepmc.org/articles/PMC15040>

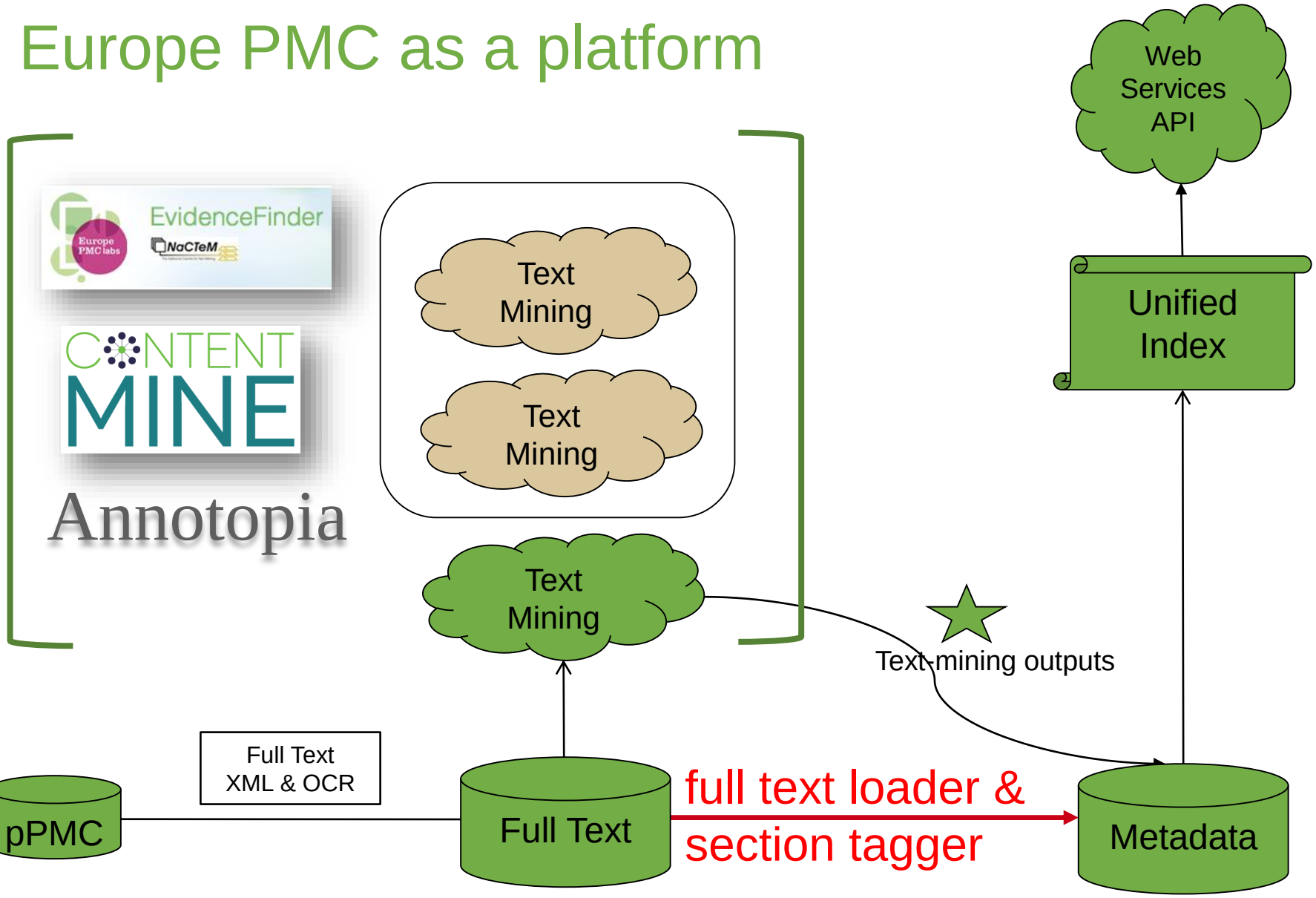
# Notes on RDF

- Running on the EBI RDF platform
- Stores 1,563,241,810 triples text-mined from ~400K OA articles
- Provides
  - For each article, all the annotations linking to ontologies/databases
  - With contexts
    - Sentences
    - Section information

# Use cases

- Show all the articles where a PDB accession number 3NSS is mentioned.
- Show all the annotations, each with its label, in PMC3382907.
- Show all the articles where inflammatory bowel disease (C0021390) is mentioned.

# Europe PMC as a platform



# Europe PMC REST Web services

- 110 indexed fields
- everything you see in the website except non-OA full text
- PMIDs and PMCIDs can be used interchangeably
- XML or JSON outputs (not full text)
- About 50 million requests per month

# Using text mining to build ontologies

- Mining novel disease-phenotype associations from EuropePMC abstracts using Whatizit for IBD and Type 2 diabetes
- EuropePMC abstracts, MeSH annotations
- Mammalian Phenotype Ontology
- Human Phenotype Ontology
- Expert clinicians
- Output – an OWL ontology with novel phenotype-disease associations
- Recall – 60%
- http://phenoday2015.bio-lark.org/pdf/6.pdf

# Centre for Therapeutic Target Validation (CTTV)

- aims to provide evidence on the biological validity of therapeutic targets … using genome-scale experiments and analysis.
- Biological impact assessment (assays to expand evidence
- Pre-competitive
- Arsenal of approaches … including text mining

# It doesn't have to be fancy

"For example, the CARD15 gene product, NOD2, influences the development of the adaptive immune response [11,12] and functional variants of the gene predispose to the inflammatory bowel disease (IBD), Crohn disease (Table 1)."

| Disease | Protein | No. of articles |
|---|---|---|
| IBD | TNF | 491 |
| IBD | IL-10 | 125 |
| IBD | CD4 | 122 |
| IBD | IL-17 | 107 |

- Appearance in title, abstract, full text
- Appearance in section e.g. introduction, conclusion, results, figure
- Article type (e.g. review)
- Publication date profile
- Self citation?……

# Patents and Chemistry

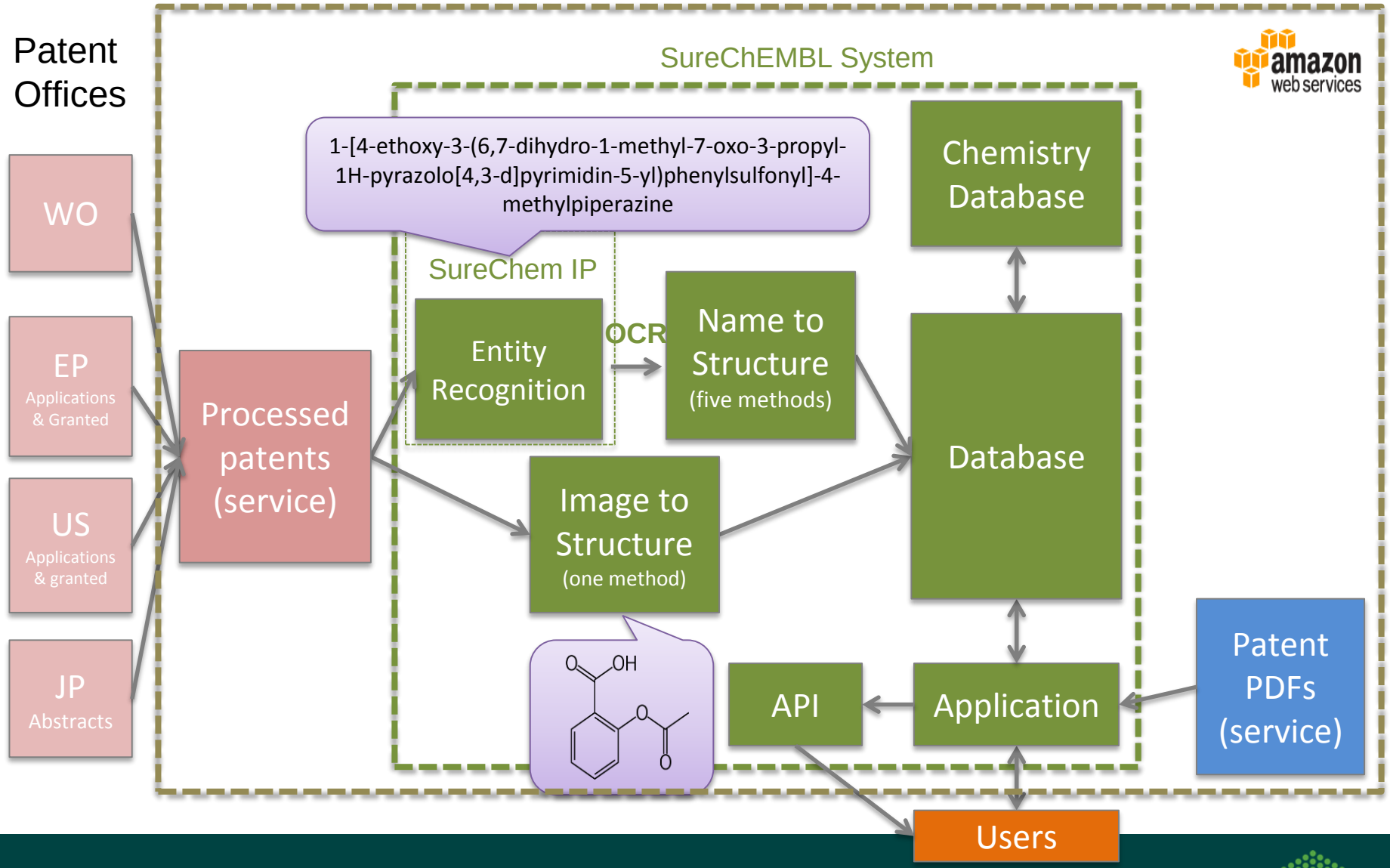
# SureChembl: Why look at patent documents?

- Patent filing and searching
  - Legal, financial and commercial incentives & interests
  - Prior art, novelty, freedom to operate searches
  - Competitive intelligence
- Unprecedented wealth of knowledge
  - Most of knowledge will never be disclosed anywhere else
  - Average lag of 2-3 years between patent document and journal publication disclosure for chemistry

# From SureChem to SureChEMBL

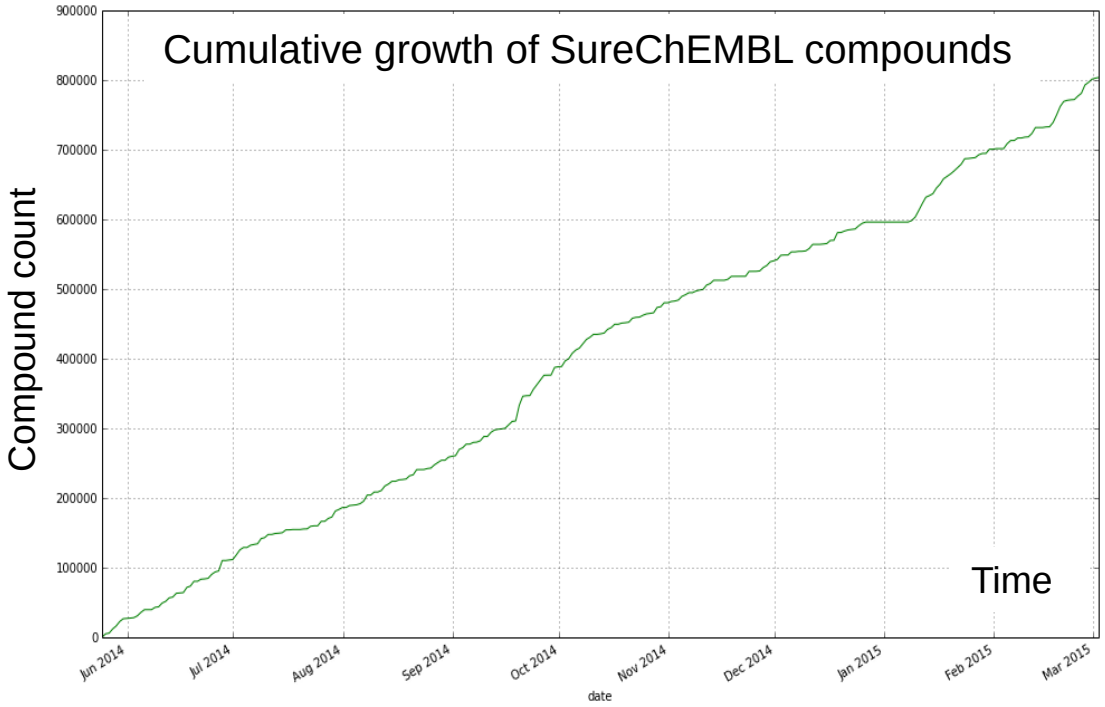
- Digital Science/Macmillan donated SureChem to EMBL-EBI
  - SureChem: commercial patent chemistry mining product
- Wellcome Trust funds further development
- EMBL-EBI provides an on-going, live service
  - Full functionality *freely* available to everyone
  - Query, view and export chemistry from patents
  - Complemented with biological annotations

# SureChEMBL data processing



# Data growth

- ~80K *novel* compounds every month
- ~800K *novel* compounds since EBI took over
- 2–7 days for a published patent to be chemically annotated and searchable in SureChEMBL



# Compound-patent map

- Flat file with
  - Compound, global frequency, document, section, section frequency, publication date
  - Back file
    - 187,958,584 unique patent-compound pairs
    - 14,076,090 unique compound IDs
    - 3,585,233 EP, JP, WO and US patent docs
    - 1960-2014
  - Quarterly incremental updates
  - Q1 2015 is also now available on the FTP

http://chembl.blogspot.co.uk/2015/03/the-surechembl-map-fi

# EMBL-EBI chemistry resources

RDF and REST API interfaces

| Atlas | PDBe | ChEBI | ChEMBL | SureChEMBL | 3rd Party Data |
|---|---|---|---|---|---|
| Ligand induced transcript response | Ligand structures from protein complexes | Nomenclature of primary and secondary metabolites. Chemical Ontology | Bioactivity data from literature and depositions | Chemical structures from patent literature | ZINC, PubChem, ThomsonPharma DOTF, IUPHAR, DrugBank, KEGG, NIH NCC, eMolecules, FDA SRS, PharmGKB, Selleck, …. |
| 750 | 15K | 24K | 1.5M | **16M** | ~65M |

UniChem – InChI-based chemical resolver (full + relaxed 'lenses') >90M

REST API Interface - https://www.ebi.ac.uk/unichem/

# UniChem


- Home / Search
- Web Services
- Connectivity Search
- Sources
- *General Info...*
  - Background
  - Getting in touch
  - FAQ
  - Downloads
  - Connectivity Info
- *Other*
- Analysis.
  - Top Level Stats
  - Structures by Source
- *Overlaps...*
  - FULIK
  - FIKHB
  - SCFIB




## Query Results...

Search terms and Sorted-by columns are highlighted.
For clarity, Standard InChIs are omitted, but may be toggled on/off.
Use the drop downs in the table footer to filter by individual columns.

Show 15 entries  Apply filter: ...to whole table

| src_id | Source Name | src_compound_id | Currently Assigned | LR * | UCI ** | Standard InChIKey |
|---|---|---|---|---|---|---|
| 1 | chembl | CHEMBL1093743 | Yes | | 570560 | GKJZMAHZJGSBKD-NMMTYZSQSA-N |
| 2 | drugbank | DB04746 | Yes | | 570560 | GKJZMAHZJGSBKD-NMMTYZSQSA-N |
| 3 | pdb | ODD | Yes | | 570560 | GKJZMAHZJGSBKD-NMMTYZSQSA-N |
| 7 | chebi | 44526 | Yes | | 570560 | GKJZMAHZJGSBKD-NMMTYZSQSA-N |
| 10 | emolecules | 26755276 | Yes | | 570560 | GKJZMAHZJGSBKD-NMMTYZSQSA-N |
| 14 | fdasrs | N151ZM4M27 | Yes | | 570560 | GKJZMAHZJGSBKD-NMMTYZSQSA-N |
| 15 | surechembl | SCHEMBL1810737 | Yes | | 570560 | GKJZMAHZJGSBKD-NMMTYZSQSA-N |
| 21 | pubchem_tpharma | 14799657 | Yes | | 570560 | GKJZMAHZJGSBKD-NMMTYZSQSA-N |
| 22 | pubchem | 5282800 | Yes | | 570560 | GKJZMAHZJGSBKD-NMMTYZSQSA-N |
| 26 | actor | 2420-56-6 | Yes | | 570560 | GKJZMAHZJGSBKD-NMMTYZSQSA-N |
| 29 | nikkaji | J604.212K | Yes | | 570560 | GKJZMAHZJGSBKD-NMMTYZSQSA-N |
| 31 | bindingdb | 50394662 | Yes | | 570560 | GKJZMAHZJGSBKD-NMMTYZSQSA-N |
| src_id | | src_compound_id | | | | |

Showing 1 to 12 of 12 entries  First Previous 1 Next Last

**Footnotes.**
* 'LR' = 'Last Release when Assignment was Current'.
** 'UCI' = 'UniChem Identifier'.
For an explanation of 'Assignments' click here.

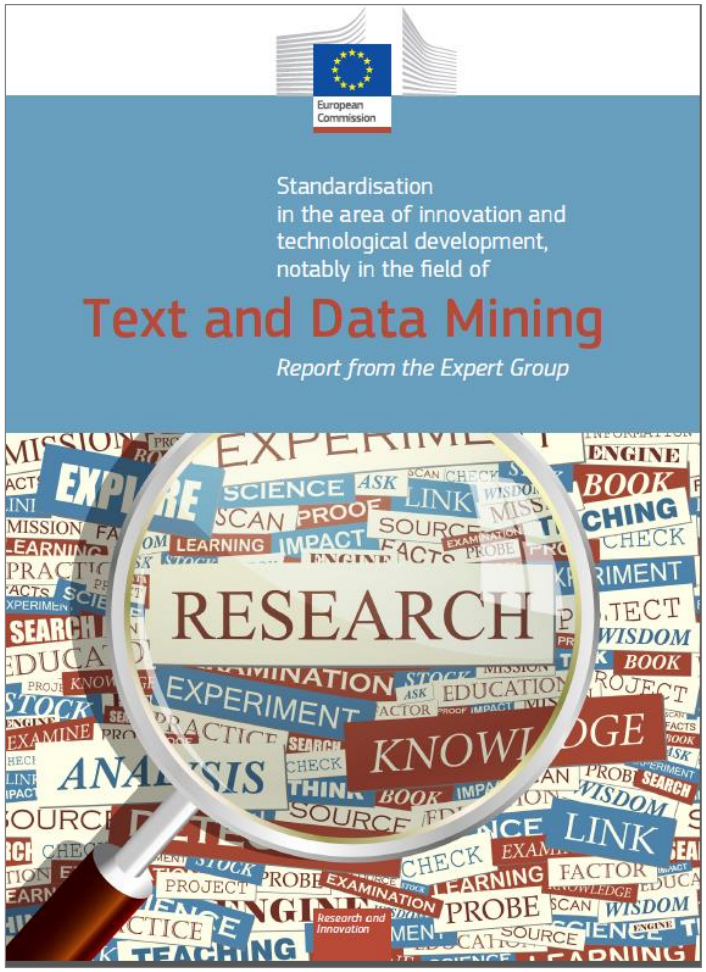
**Back to UniChem Home and Query page.**

BioMedBridges  EU-OPENSCREEN Chemical keys for life's locks

# Reuse: Text and Data Mining (TDM)

- Text mining treats articles as **big data**
- UK: changes to legislation
- Europe: addressing barriers to TDM → 2% (€5.3 billion) increase in the real value of research output produced by the EU research budget.

The Hague DECLARATION



# Europe PMC is funded by:



























## Delivered by: