Interoperability of large scale image data sets with ontologies

CrEDIBLE wokshop

10 October 2014, Sophia Antipolis

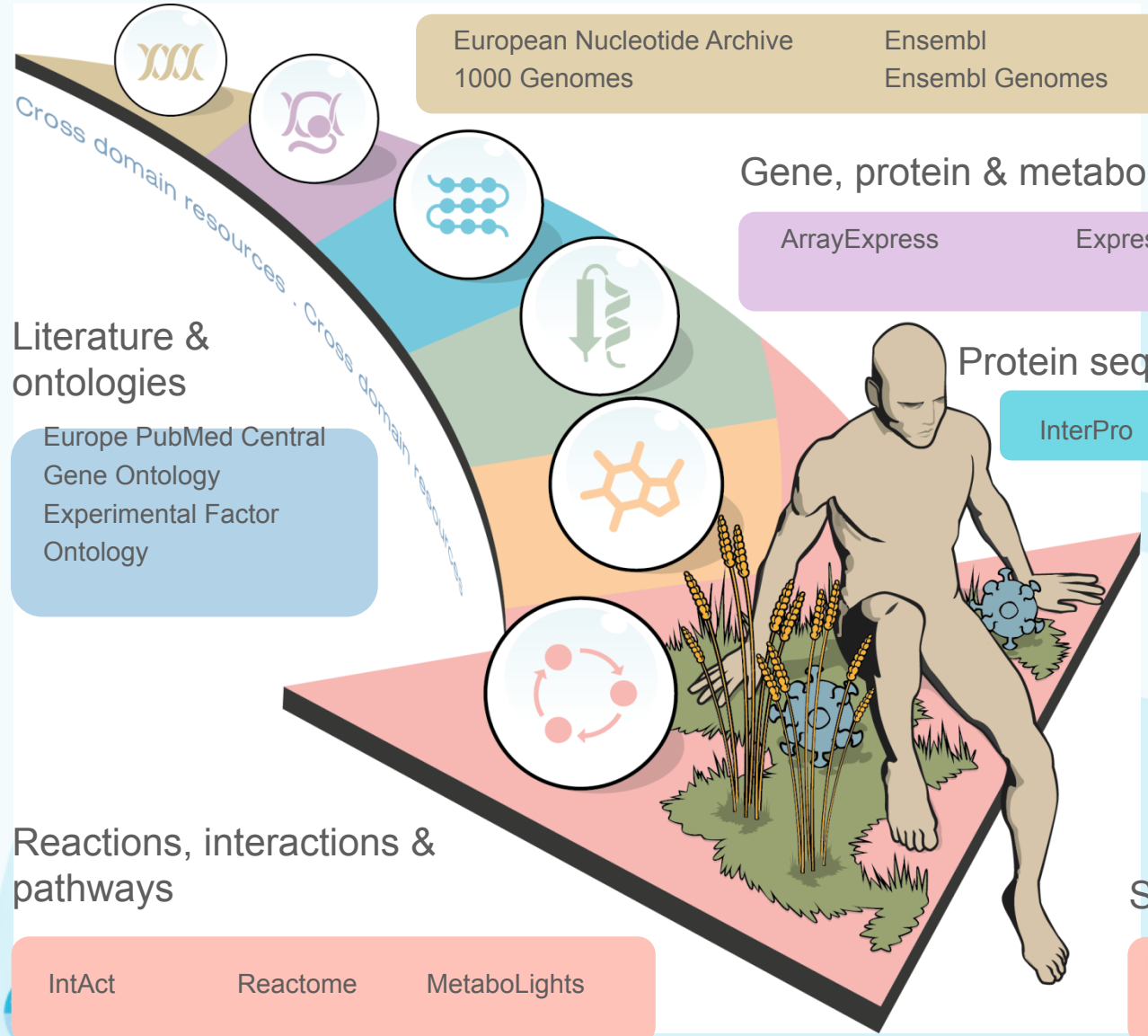
Simon Jupp
Samples Phenotypes and Ontologies Team
EMBL-European Bioinformatics Institute

EMBL-EBI

Data resources at EMBL-EBI

Genes, genomes & variation

European Nucleotide Archive
1000 Genomes
Ensembl
Ensembl Genomes
European Genome-phenome Archive
Metagenomics portal



Gene, protein & metabolite expression

ArrayExpress
Expression Atlas
Metabolights
PRIDE

Literature & ontologies

Europe PubMed Central
Gene Ontology
Experimental Factor Ontology

Protein sequences, families & motifs

InterPro
Pfam
UniProt

Molecular structures

Protein Data Bank in Europe
Electron Microscopy Data Bank

Chemical biology

ChEMBL
ChEBI

Reactions, interactions & pathways

IntAct
Reactome
MetaboLights

Systems

BioModels
Enzyme Portal
BioSamples

Sample description with semantic markup

Export Experiment Design as Tab-Delimited file

Assay	Array	cell type	Compound treatment	organism
GSM589566	A-AFFY-141	human umbilical vein endothelial cell	valproic acid	Homo sapiens
GSM589557	A-AFFY-141	human umbilical vein endothelial cell	none	Homo sapiens
GSM589564	A-AFFY-141	human umbilical vein endothelial cell	valproic acid	Homo sapiens
GSM589558	A-AFFY-141	human umbilical vein endothelial cell	valproic acid	Homo sapiens
GSM589559	A-AFFY-141	human umbilical vein endothelial cell	none	Homo sapiens
GSM589563	A-AFFY-141	human umbilical vein endothelial cell	none	Homo sapiens
GSM589561	A-AFFY-141	human umbilical vein endothelial cell	none	Homo sapiens
GSM589565	A-AFFY-141	human umbilical vein endothelial cell	none	Homo sapiens
GSM589560	A-AFFY-141	human umbilical vein endothelial cell	valproic acid	Homo sapiens
GSM589562	A-AFFY-141	human umbilical vein endothelial cell	valproic acid	Homo sapiens

CL:CL_0000071 (blood vessel endothelial cell)

obo:CHEBI_39867 (valproic acid)

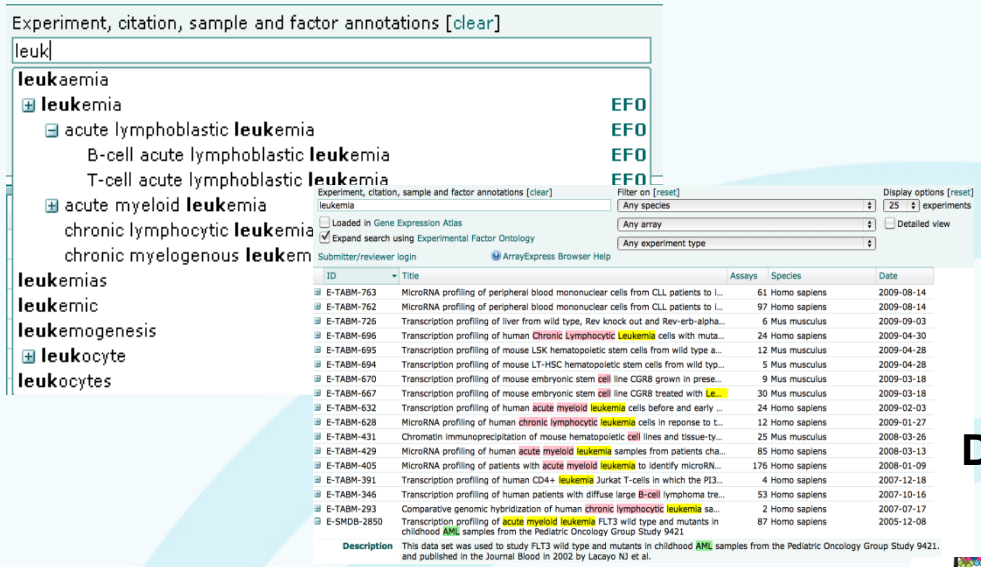
NCBITaxon:NCBITaxon_9606 (Homo Sapiens)

BioMedBridges

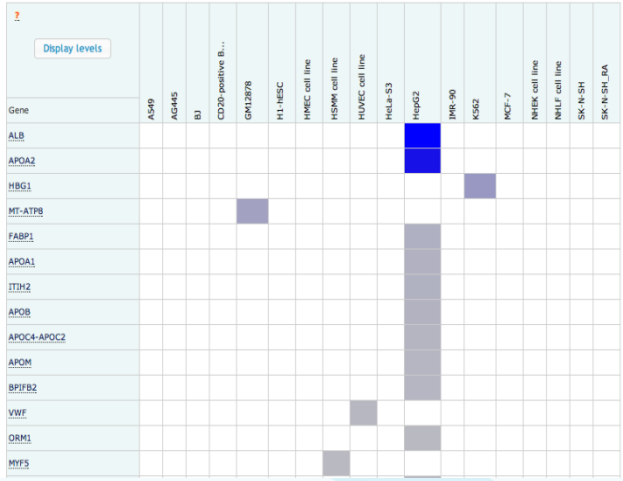
SEVENTH FRAMEWORK PROGRAMME

Ontologies add value

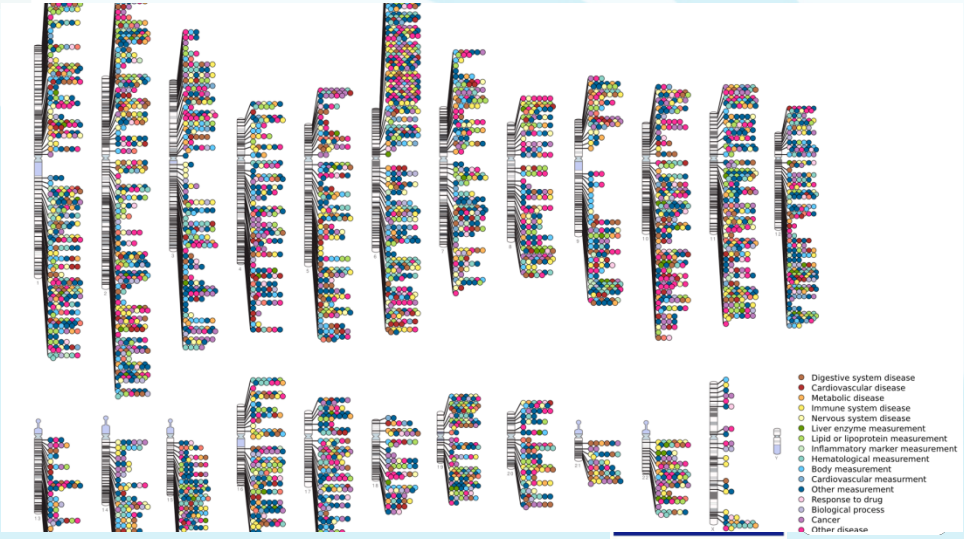
Smarter searching



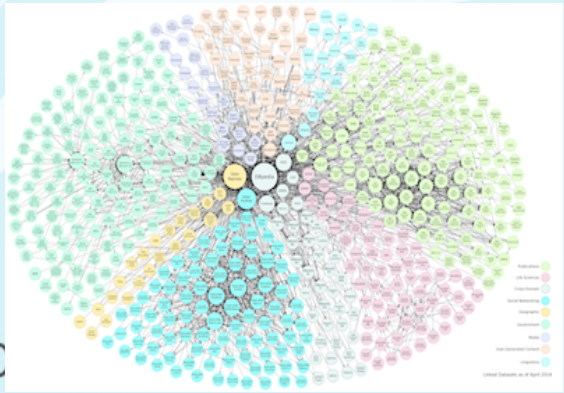
Data analysis



Data visualisation

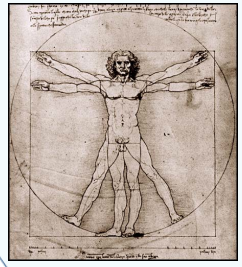


Data integration



Bio

BioMedBridges Project



EU-OPENSCREEN
Chemical keys for life's locks

elixir

BBMRI
Biobanking and Biomolecular Resources Research Infrastructure

instruct
Integrating Biology

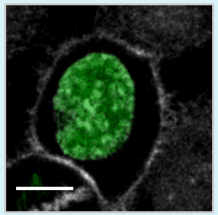
erinha
European Research Infrastructure on Highly Pathogenic Agents

EMBRC
EUROPEAN MARINE BIOLOGICAL RESOURCE CENTRE

ECRIN
EUROPEAN CLINICAL RESEARCH INFRASTRUCTURES NETWORK

EATRIS
European Infrastructure for Translational Medicine

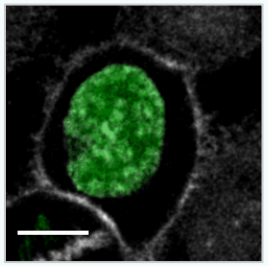
EURO-BIOIMAGING



BioMedBridges

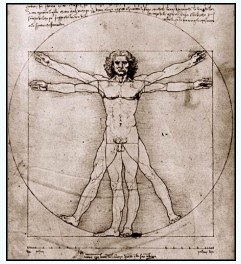
SEVENTH FRAMEWORK PROGRAMME

Scientific problem



Cell

Mouse



Human

Genome

- ✓ Cellular phenotype
- ✓ Genetic information
- ✓ Molecular mechanism

- ✓ Tissue phenotype
- ✓ Genetic information

- ✓ Tissue phenotype

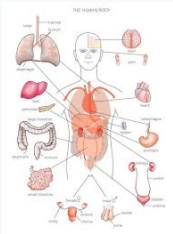
Imaging → **Imaging** → **Imaging**

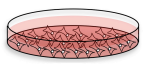
By linking these three different types of data sets, we can better understand diseases, predict novel drug targets and biomarkers

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

To compare and integrate image data we need interoperable standards

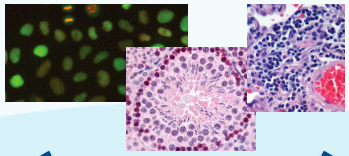




Sample

Assay

Images



Volocity MVD2
Olympus OIB
JPEG
OME-TIFF
PNG
Zeiss LSM
NDP
Leica LIF
DeltaVision DV
HDF5
Olympus OIF

Different file formats

No consistent phenotype annotation/ontology

Different image metadata

Automated comparative analysis of image data sets was impossible

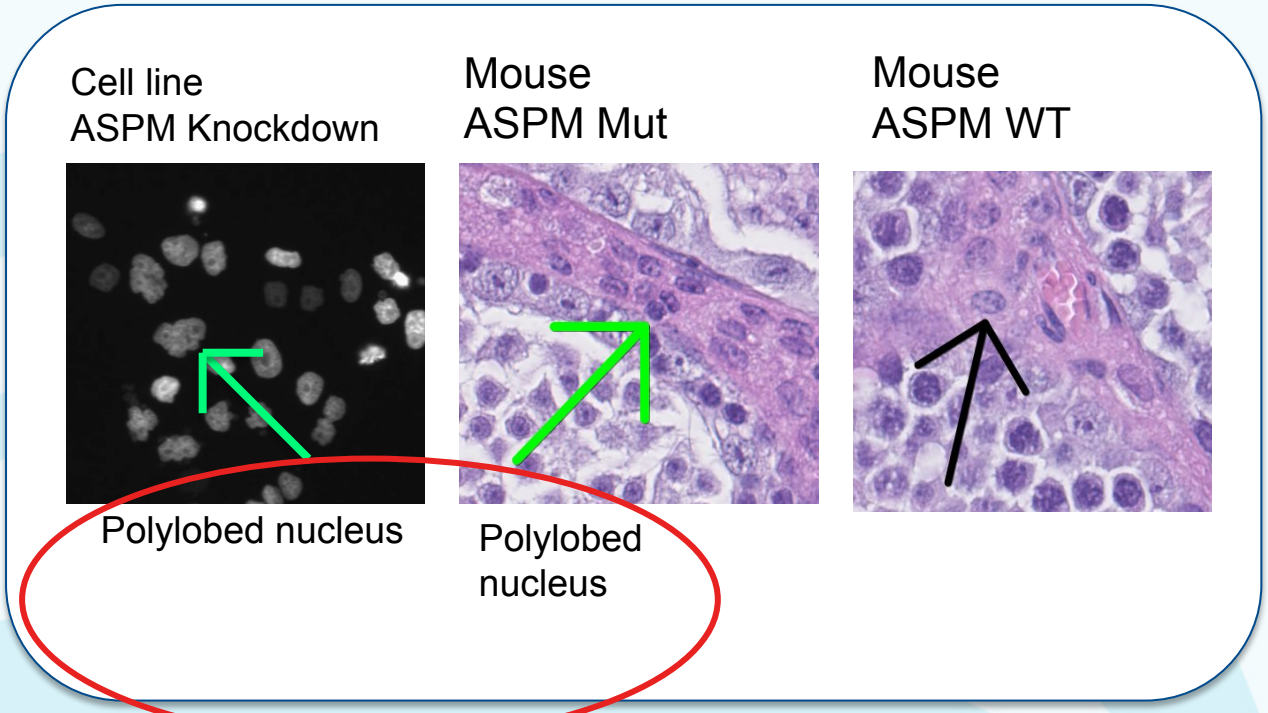
BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

Correlative analysis and biomarker prediction

Promising gene candidates from cellular screens

- MLL3
- PAPPA
- SF3B1
- PRPF8
- CENPE
- CIT
- **ASPM**
- ESPL1
- DYNC1H1
- ASCC3
- KIF4A



Mouse and human tissue WP6 partners are looking for and/or generating the data relevant to these genes to be used for analysis.

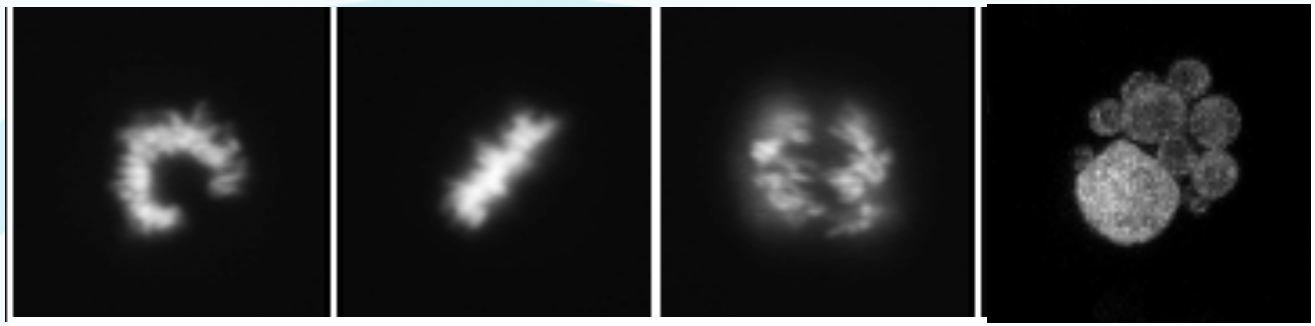
BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

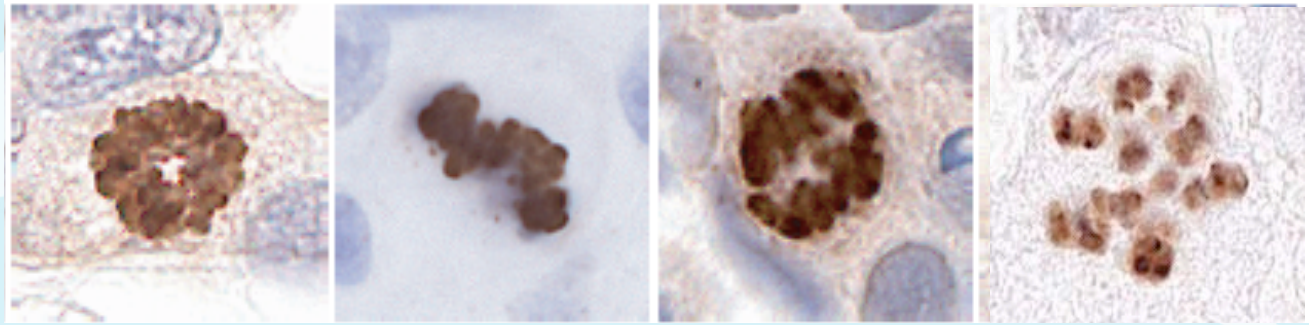
Matching phenotypes at different scales

Prometaphase | Metaphase | Anaphase | Graped micronucleus

Cell line – gene knockdown



Human cancer tissue



State of the art: finding a match by chance

Use cases

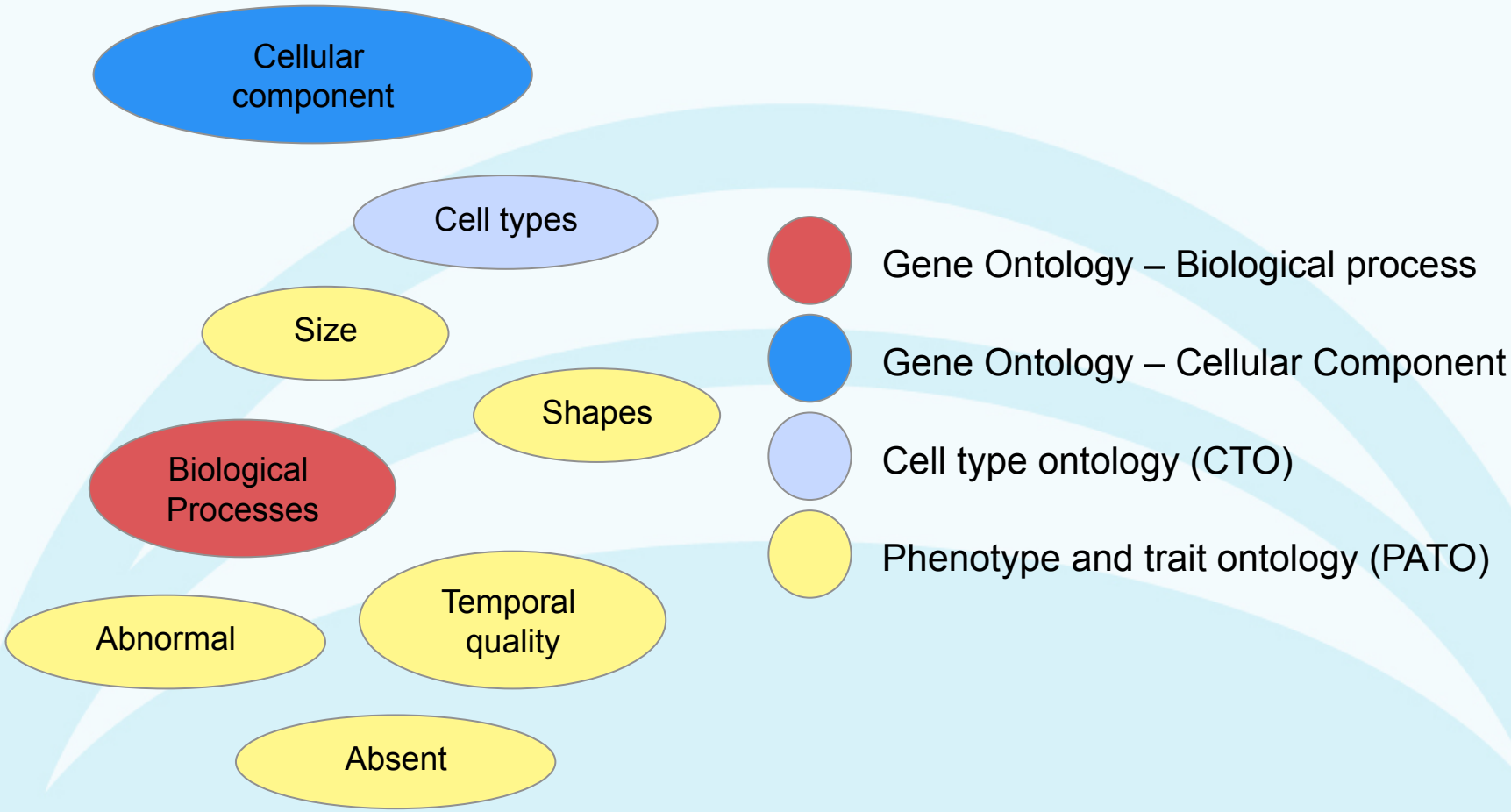
- Defined classes: *"Abnormal cellular membrane phenotypes"* includes all parts of the cellular membrane
- Real queries
  - Genes / siRNAs / Images having a mitotic phenotype in all screens or in one particular screen.
  - Genes / siRNAs / Images having a cytokinesis phenotype in all screens or in one particular screen.
  - Genes / siRNAs / Images having a mitotic phenotype, but no cell death.
  - Genes / siRNAs / Images having a mitotic phenotype, but no problem in Prophase.
  - Genes / siRNAs / Images having a mitotic phenotype followed by cytokinesis defects.
  - Genes / siRNAs / Images having a mitotic phenotype before 30 hours in all screens or in one particular screen.
  - Genes / siRNAs / Images having a mitotic phenotype restricted to human and Drosophila.

Existing ontologies are not enough

- Existing ontologies either lack coverage or are incomplete to describe cellular scale phenotypes
- No species neutral ontology for cellular phenotypes
- Such ontology is needed for data interoperability

➢ **WP6 developed the**
Cellular Microscopy Phenotype Ontology (CMPO)

Cellular phenotypes: entities, processes and qualities



BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

Building a phenotype ontology

Composing a phenotype description

Entity a bearer of some quality	+	Quality characteristic of the entity

Examples:

- Phenotype: "*Large nucleus*"
  - Entity: nucleus (GO_000xxxx)
  - Quality: large (PATO_000xxxx)
- Phenotype: "*Cells stuck in metaphase due to metaphase arrest*"
  - Entity: mitotic metaphase (GO_0000089)
  - Quality: arrested (PATO_0000297)

Cellular Microscopy Phenotype Ontology (CMPO)

- Annotate the available data using EQ based ontology annotations (post composition)
- Translate EQs into OWL axioms following OBO style “part-of some” pattern
- Create new terms via post composition and import into the CMPO ontology (assign labels, definitions, provenance)

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

Composing ontology terms from annotations in OWL

- More expressive, faster reasoners (ELK)
- Basic modeling pattern *(has_part some (<Quality> and inheres_in some <Entity>))*

Original phenotype: metaphase arrest

↓

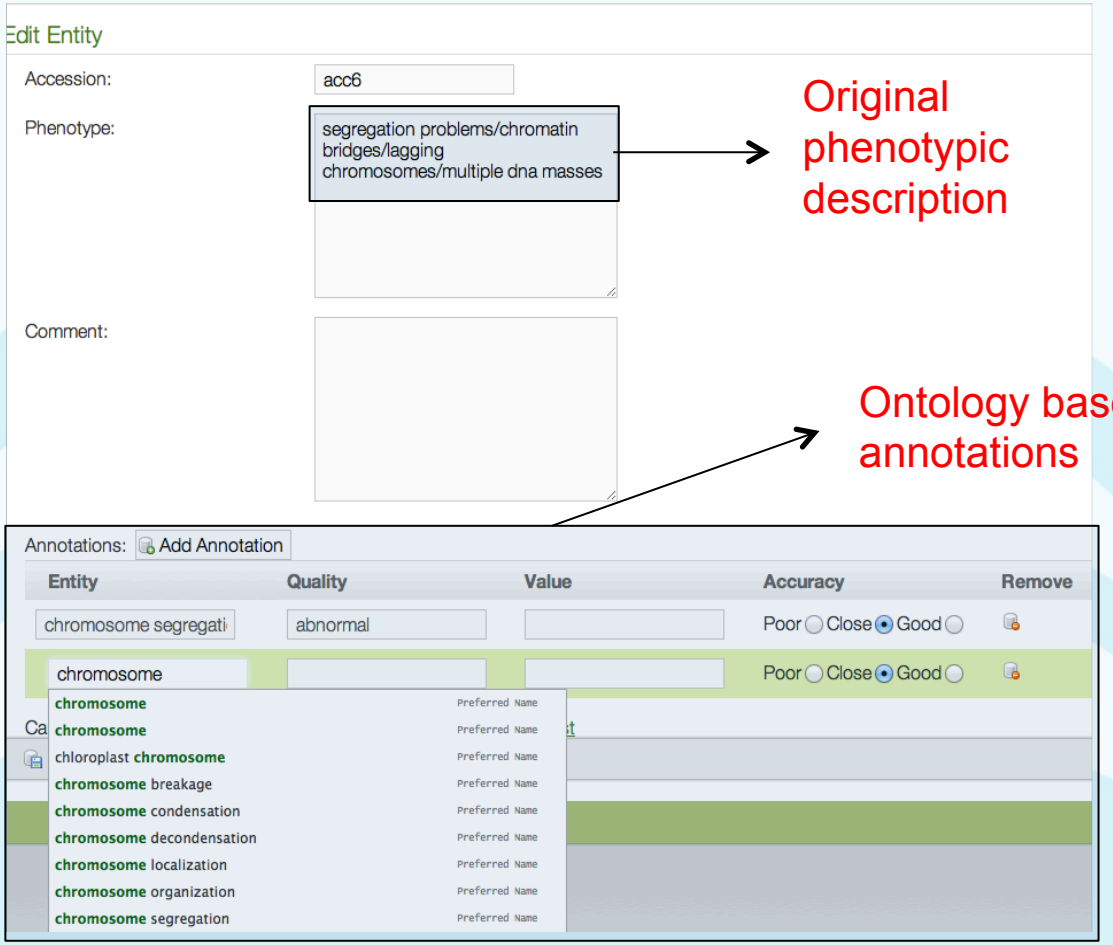
EQ annotation :
mitotic metaphase (GO_0000089)
arrested (PATO_0000297)

↓

Translation to OWL class description:
CMPO_0000305 **equivalentTo**
has_part **some** (‘arrested’
and (*inheres_in* **some** 'mitotic metaphase'))

Enabling standardised data generation

Phenotator: user-friendly ontology annotation of image data

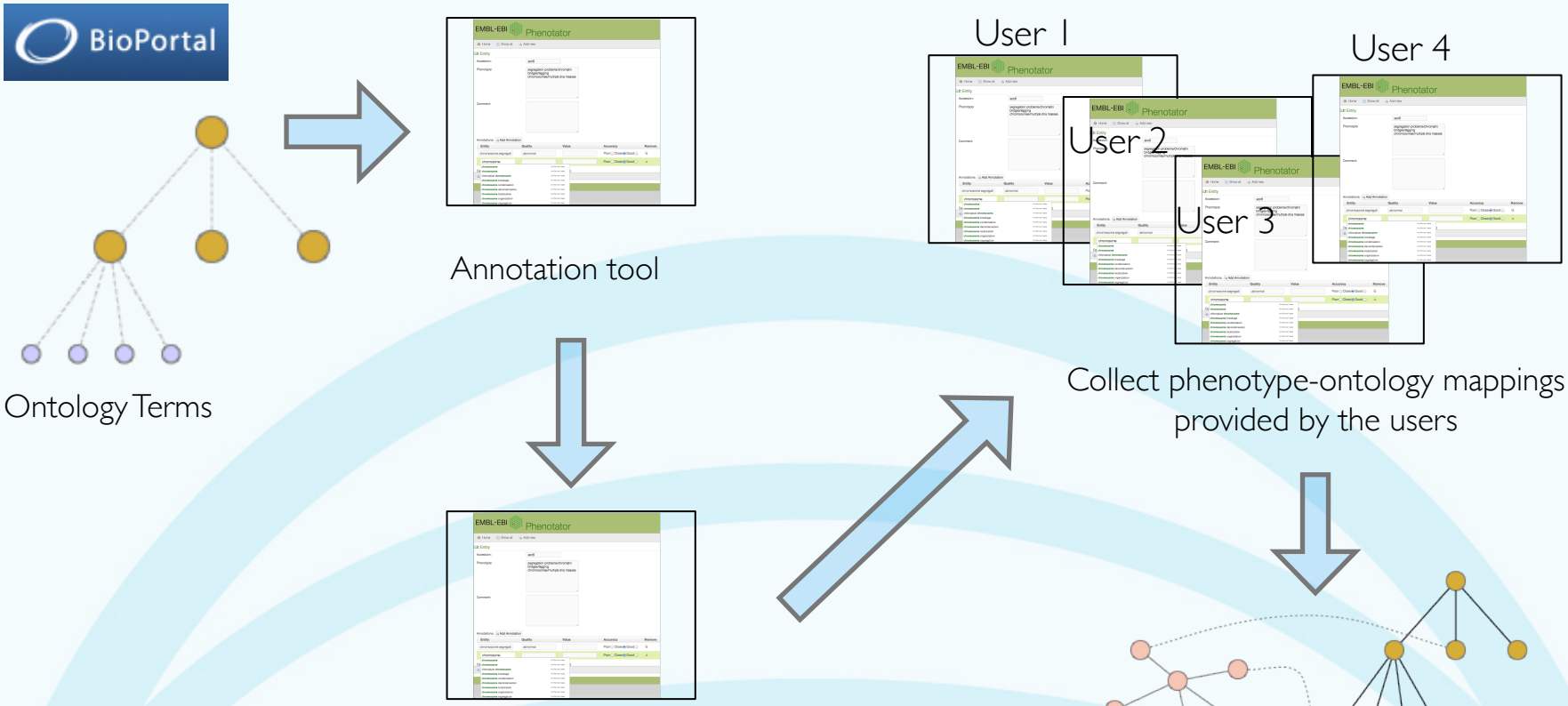


http://wwwdev.ebi.ac.uk/fgpt/phenotator/

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

Ontology building using Phenotator



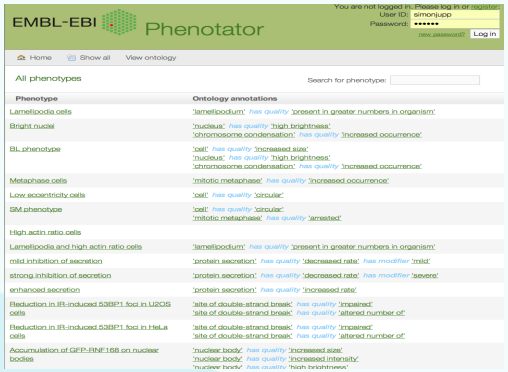
1. Distribute tool to consortium members for phenotype annotation
2. Workshop on ontology development with WP6 partners
3. One to one sessions with data producers

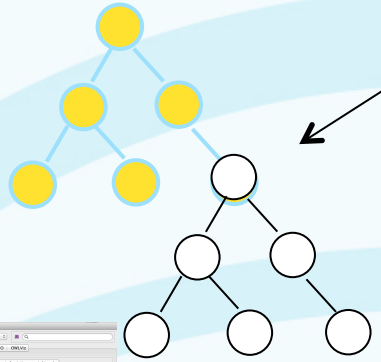
BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

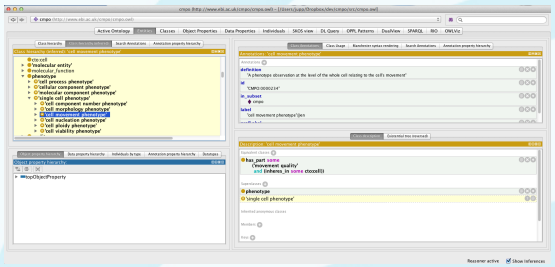
Building the ontology

CMPO upper level species neutral ontology
(cell process phenotypes,
cellular component phenotypes,
whole cell phenotypes,
cell population)

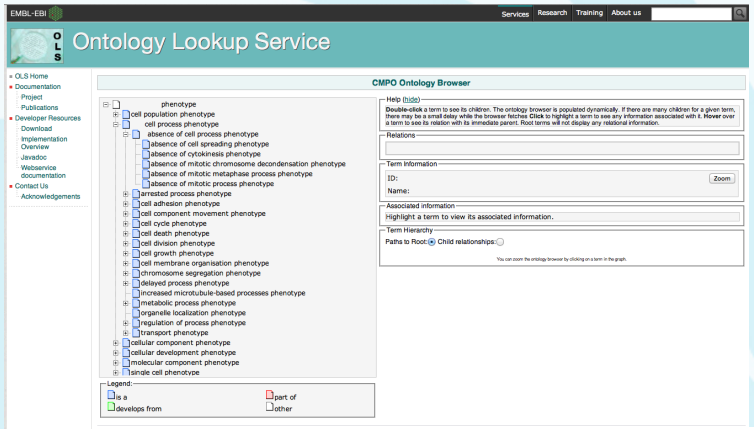




Convert annotation to OWL axioms and owl:import and classify under upper level CMPO



Ontology editor merges terms,
adds definitions,

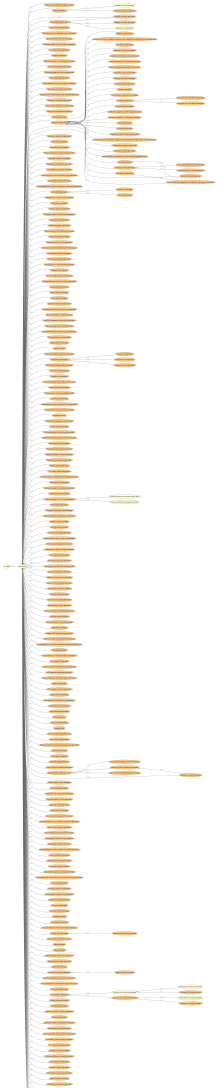


Biologists verify new classification

BioMedBridges

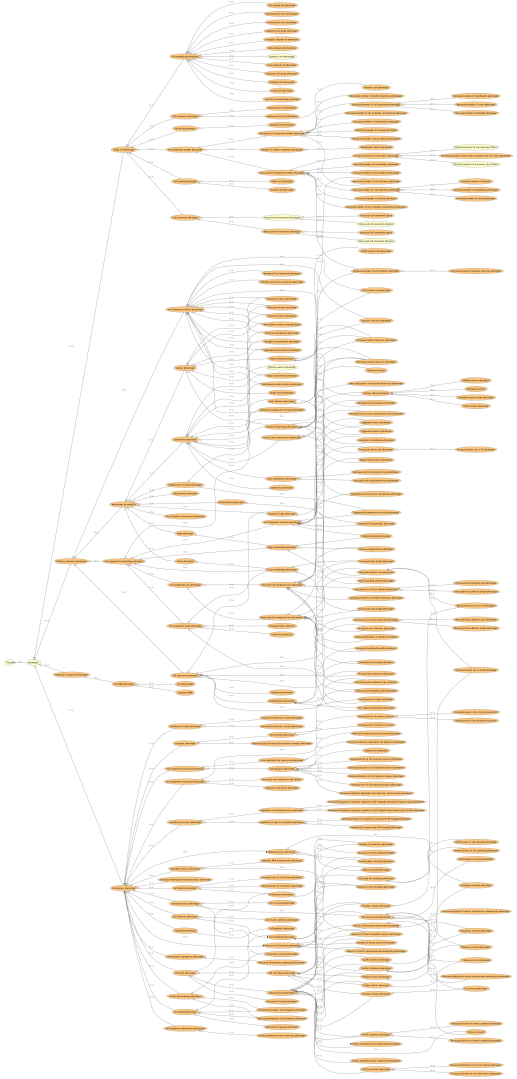
SEVENTH FRAMEWORK PROGRAMME

Exploit GO and PATO to automatically construct CMPO hierarchy



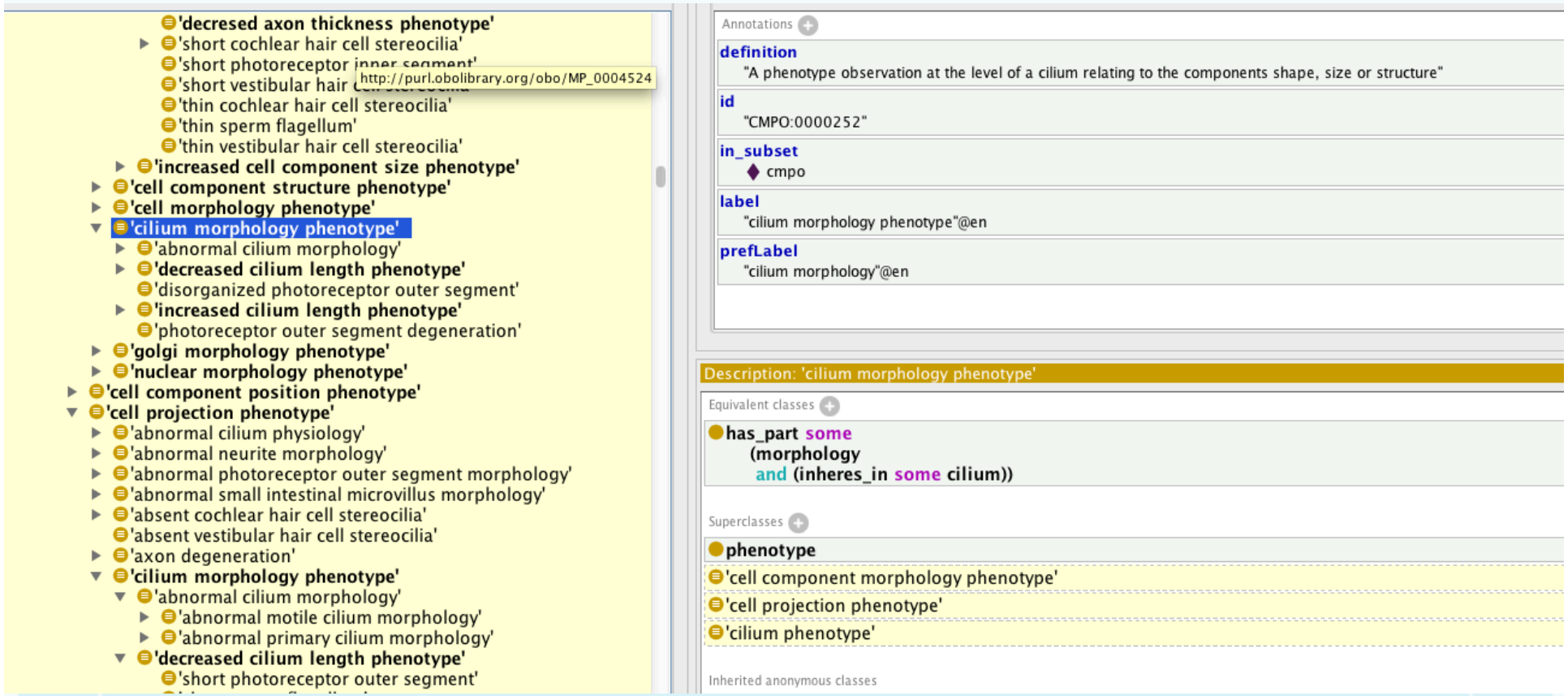
Relatively flat asserted view

OWL reasoner



Inferred polyhierarchy

Inferring equivalence across species (MP, FYPO and CMPO)

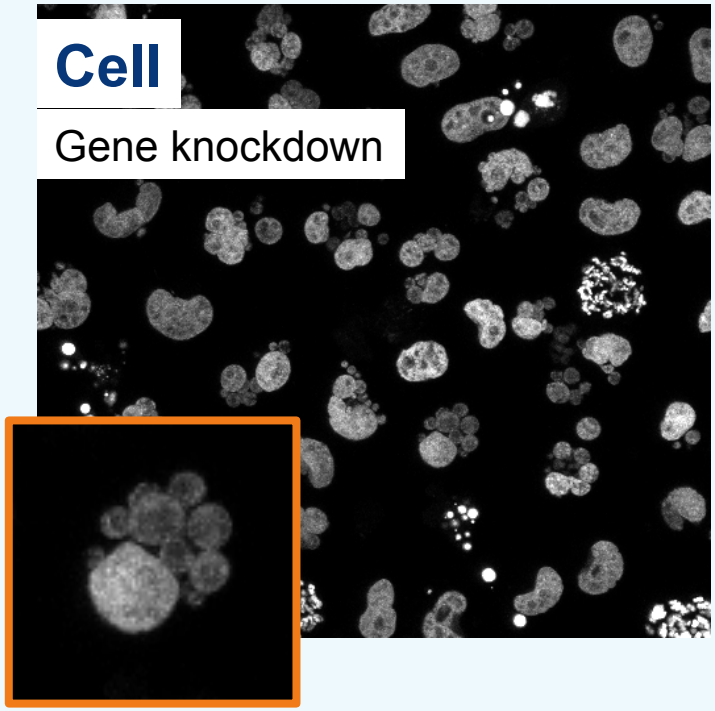


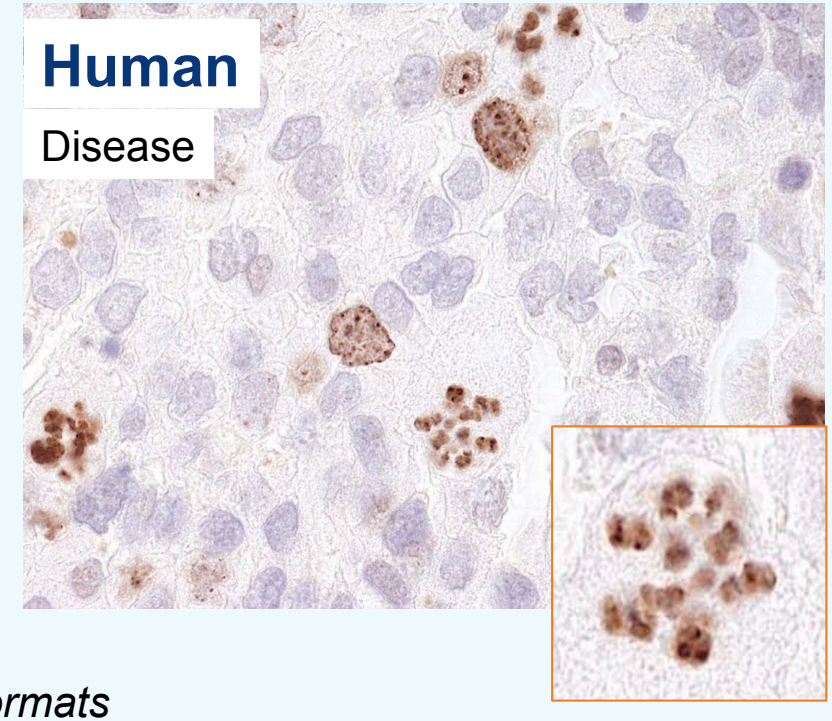
BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

Automated ontology mapping with

elongated cells	elongated cell phenotype	Automatic	CMPO_0000077	SysMicro
lamellipodia + high actin ratio cells	more lamellipodia cells, increased number of actin filament phenotype	Automatic	CMPO_0000083, CMPO_0000105	SysMicro
lamellipodia cells	more lamellipodia cells	Automatic	CMPO_0000083	SysMicro
sm phenotype	metaphase arrested phenotype, round cell phenotype	Automatic	CMPO_0000305, CMPO_0000118	SysMicro
reduction in ir-induced 53bp1 foci in u2os cells	site of double-strand break phenotype, decreased number of site of double-strand break phenotype	Automatic	CMPO_0000180, CMPO_0000181	SysMicro
accumulation of gfp-rnf168 on nuclear bodies	bright nuclear body phenotype, increased nuclear body size phenotype	Automatic	CMPO_0000335, CMPO_0000126	SysMicro
intracellular retention of sh4(yes)-mcherry	decreased rate of intracellular protein transport phenotype	Automatic	CMPO_0000346	SysMicro
no nf-kb oscillation	cytoplasmic sequestering of NF-kappaB phenotype	Automatic	CMPO_0000332	SysMicro
decreased nf-kb oscillation	decreased frequency of negative regulation of NF-kappaB transcription factor activity phenotype	Automatic	CMPO_0000330	SysMicro
cell shape processes or spiky or stretchy	star shaped cell phenotype	Automatic	CMPO_0000267	SysMicro
increased number of actin puncta or dots	increased amount of punctate actin foci phenotype	Automatic	CMPO_0000291	SysMicro
increased number of zigzag actin stress fibers	increased amount of zig-zag stress fibers	Automatic	CMPO_0000299	SysMicro





Integrate file formats

Integrate metadata

Apply phenotype ontology

<u>CMPO term:</u>

graped micronucleus

CMPO_0000156

<u>CMPO term:</u>

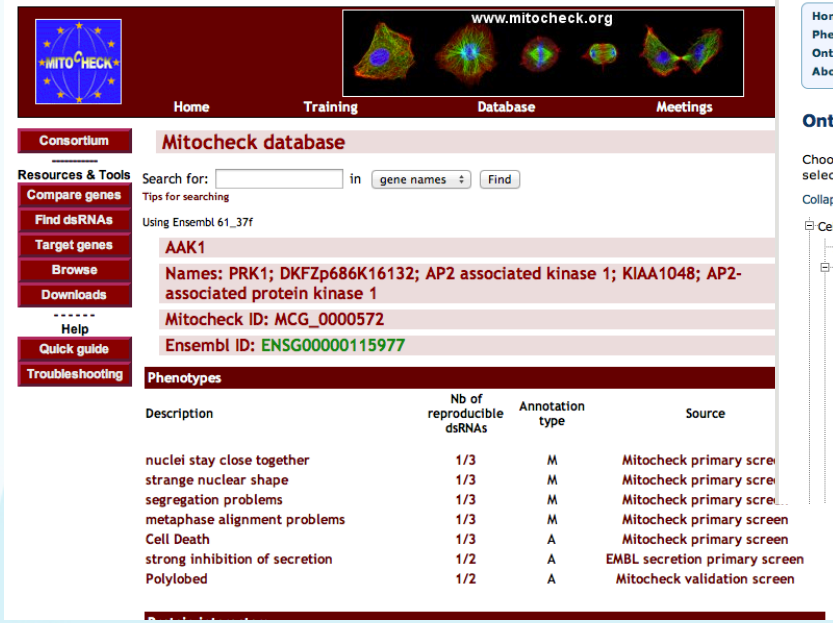
graped micronucleus

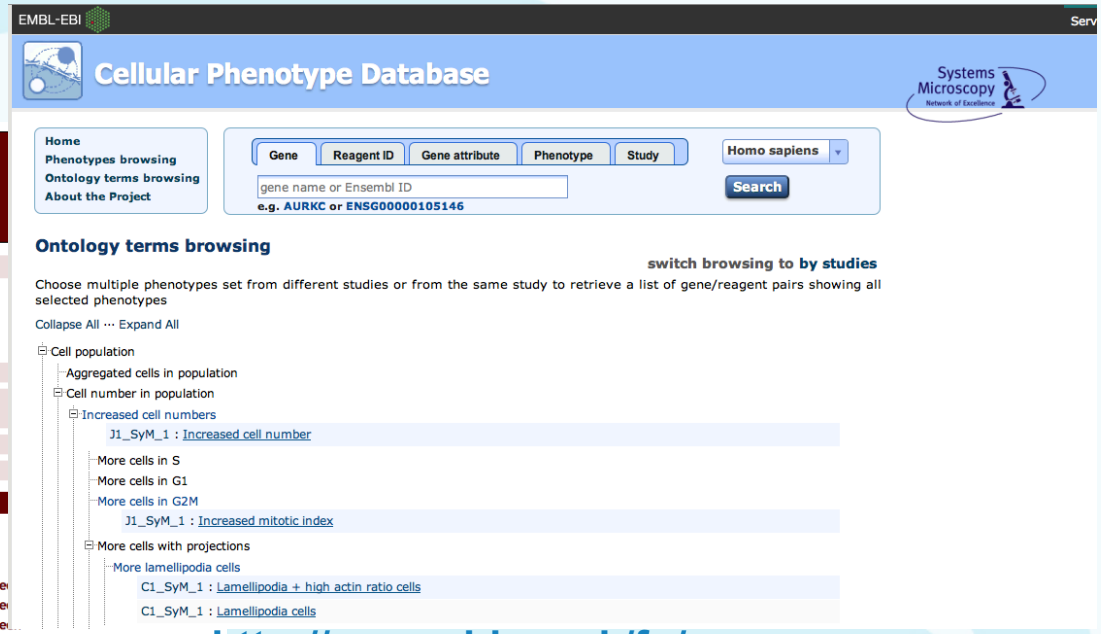
CMPO_0000156

Predict disease gene/biomarkers

Integrating CMPO

- Cellular phenotype database (EMBL-EBI), Mitosys.org (EMBL),Webmicroscope (coming soon)
- OMERO, Phenoimage share, CellCognition, CellProfiler, Knime, ImageJ, Bioconductor, GenomeRNAi





http://www.ebi.ac.uk/fg/sym

http://www.mitosys.org

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

EBI Collaborations with Image Generation Projects

- KOMP2 – adult and mouse embryonic phenotyping (WTSI, MRC Harwell)
- HIPSCI – Cellular Imaging Watt Lab of iPSCs (WTSI, EBI, Lamond)
- BioMedBridges – Cross scale image integration – EuroBioImaging, Jan Ellenberg
- EMDB – Electron Micrographs (Gerard Kleywegt)
- Virtual FlyBrain (O'Kane, Jefferis, Armstrong, David Osumi-Sutherland)
- BioStudies – Unstructured Data (Alvis Brazma)
- EBiSC – iPSC – via CellFinder – Andreas Kurtz, Charite

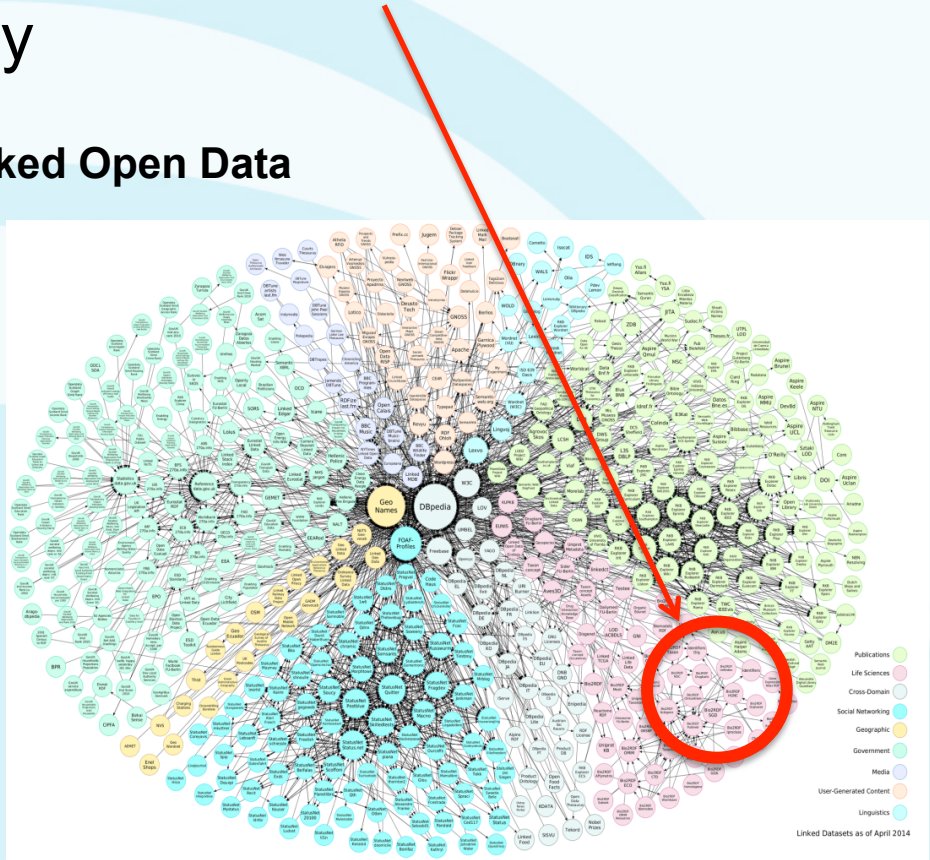
BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

Publishing biological data as Linked Open Data

- The EBI RDF platform
  - Released Nov 2013
  - Currently over 16 billion RDF triples
- Datasets updated ~ quarterly

Jupp et al (2013). **The EBI RDF Platform: Linked Open Data for the Life Sciences**. Bioinformatics.



LOD diagram August 2014

BioMedBridges

Bridging the semantic gap

- SPARQL extensions to support OWL queries
  - OWLET, Aber-OWL
- Dedicated SPARQL endpoints that can answer OWL queries

```
SERVICE <http://www.ebi.ac.uk/ols-owl/sparql> {
  GRAPH <http://www.ebi.ac.uk/cmpo> {
    ?classes ebi-ols:subclasses
            "phenotype and inheres_in some 'mitotic metapahse'"^^mosi .
 }
}
```

- Bit of a hack, need triple stores capable of EL plus nicer SPARQL syntax for writing OWL class expressions

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

Summary

- We have a pipeline for building an application ontology for cellular microscopy data
  - The biologists build define the ontology terms by annotating data
- Need for templates to guide the annotators e.g. “Increased cytoplasmic actin”
  - EQ (‘actin filament’, ‘present in greater number in organism’)
  - EQE2 (‘actin filament’, ‘localised’, ‘cytoplasm’)
  - EQE2 (‘cytosol’, ‘has extra parts of type’, ‘actin filament’),
  - EQ (‘localisation of actin to cytosol’, ‘increased rate’)
- NCBO Driving biological project grant underway to generify the phenotator approach to OWL ontologies
  - Building ontologies from Google spreadsheets

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

BMS RI partners

Euro-BioImaging

Jan Ellenberg

Tanja Ninkovic

Jean-Karim Heriche

Wolfgang Huber

Elixir

Gabriella Rustici

Simon Jupp

Infrafrontier

Frauke Neff

Philipp Gormanns

BBMRI

Johan Lundin

Mikael Lundin

Acknowledgments

- WP6 partners
- James Malone, Tony Burdett and Helen Parkinson, EMBL-EBI
- In particular, we wish to thank:
  - Anna Melidoni, Ruth Lovering and Jennifer Rohn (UCL)
  - Beate Neumann and Jean Karim Heriche (EMBL)
  - Bob Van De Water (U. Leiden)
  - Bram Herpers (OcellO)
  - Claudia Lukas (U. Copenhagen)
  - Greg Pau (Genentech)
  - Sylvia Le Dévédec (LUMC)
  - Thomas Walter (Institut Curie)
  - Wies Roosmalen (U. Twente)
  - Zvi Kam (Weizmann Institute)

SEVENTH FRAMEWORK PROGRAMME

Thank you for your attention.

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME