

# Exploiting semantic technologies to build an application ontology

James Malone PhD, Helen Parkinson PhD, Tomasz Adamusiak Phd, MD,  
Ele Holloway

EMBL-EBI



# Overview



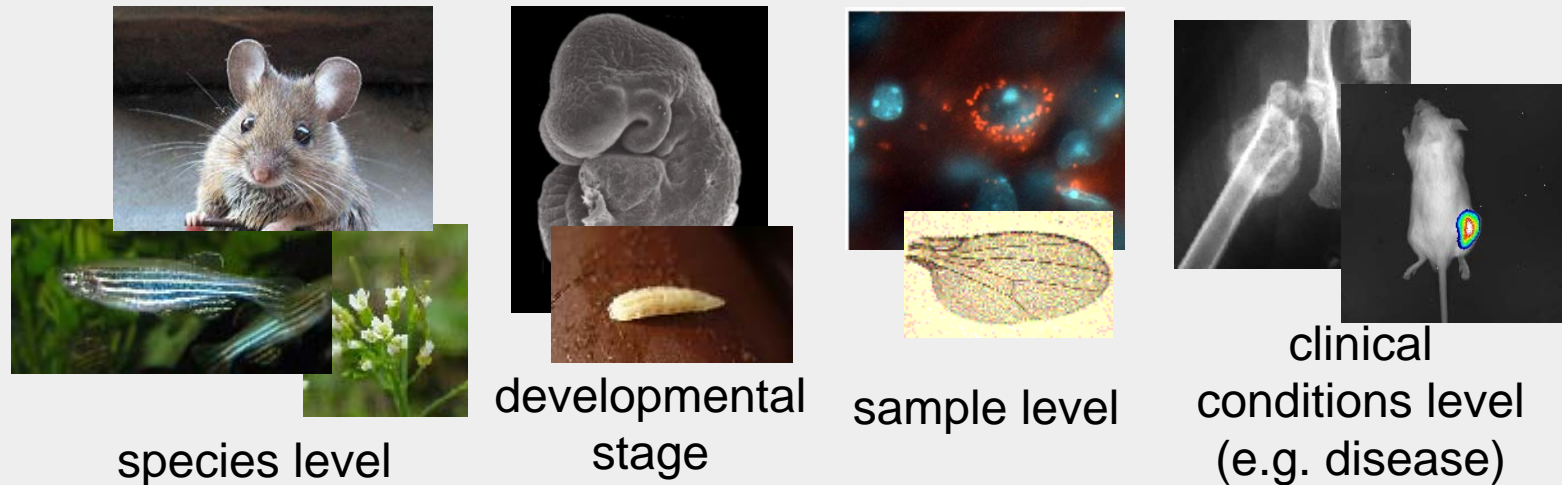
- Motivation
  - Our use cases
  - Annotating HTP experimental data
  - Integrating clinical data
- Methodology for creating the ontology
  - Semi-automated mapping and manual curation
- Current ontology usage
- Future use

# Our Use Cases

- Query support (e.g, query for 'cancer' and get also 'leukemia')
- Over-representation analysis in groups of samples (analogous to the use of GO terms in over-representation analysis in groups of genes)
- Data visualisation – e.g., presenting an ontology tree to the user of what is in the database
- Data integration by ontology terms – e.g., we assume that 'kidney' in independent studies roughly means the same, so we can count how many kidney samples we have in the database
- Intelligent template generation for different experiment types in submission or data presentation
- Summary level data
- Nonsense detection – e.g. telling us that something marked as cancer can not be marked as healthy

# Scope of Experimental Factor Ontology (EFO)

- Modelling all of the experimental factors that are currently present in the ArrayExpress repository
- Experimental factors are variable aspects of an experiment design which can be used to describe an experiment
- Scope is primarily determined by data currently held in ArrayExpress



# 'Experimental Factors'

**E-MEXP-114** Transcription profiling of hypothalamus, liver, kidney, ovaries and testis from male and female humans and mice

Gene Expression Profile for (Acsm2)

ORGANISM ORGANISMPART SEX

Gene Properties [ expand ]

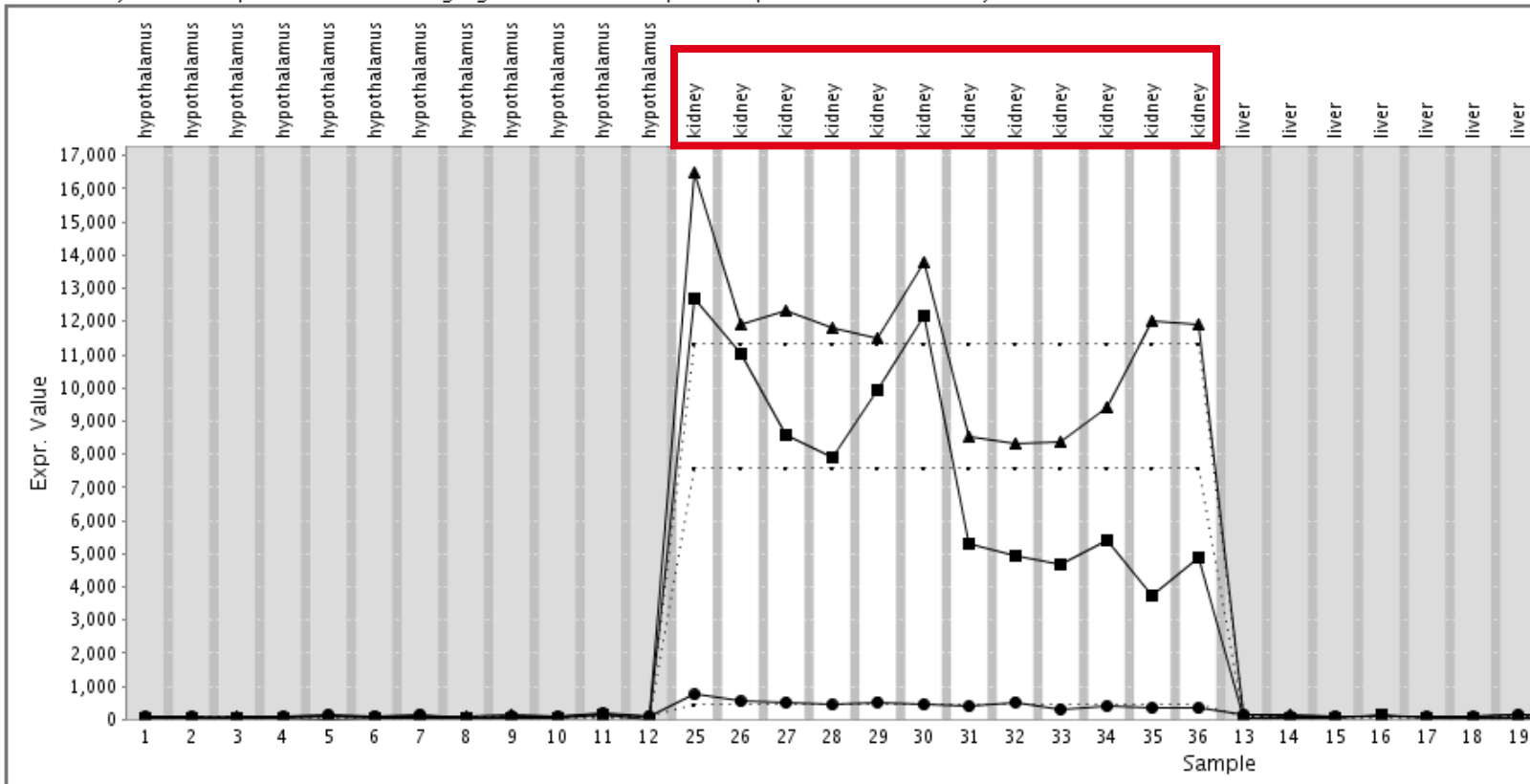
**Gene name**

**Acsm2**  
 1427223\_a\_at  
 1427224\_a\_at  
 1456190\_a\_at  
 BC031140

similarity search

- Gene Ontology: butyrate-CoA ligase activity, catalytic activity, metabolic process

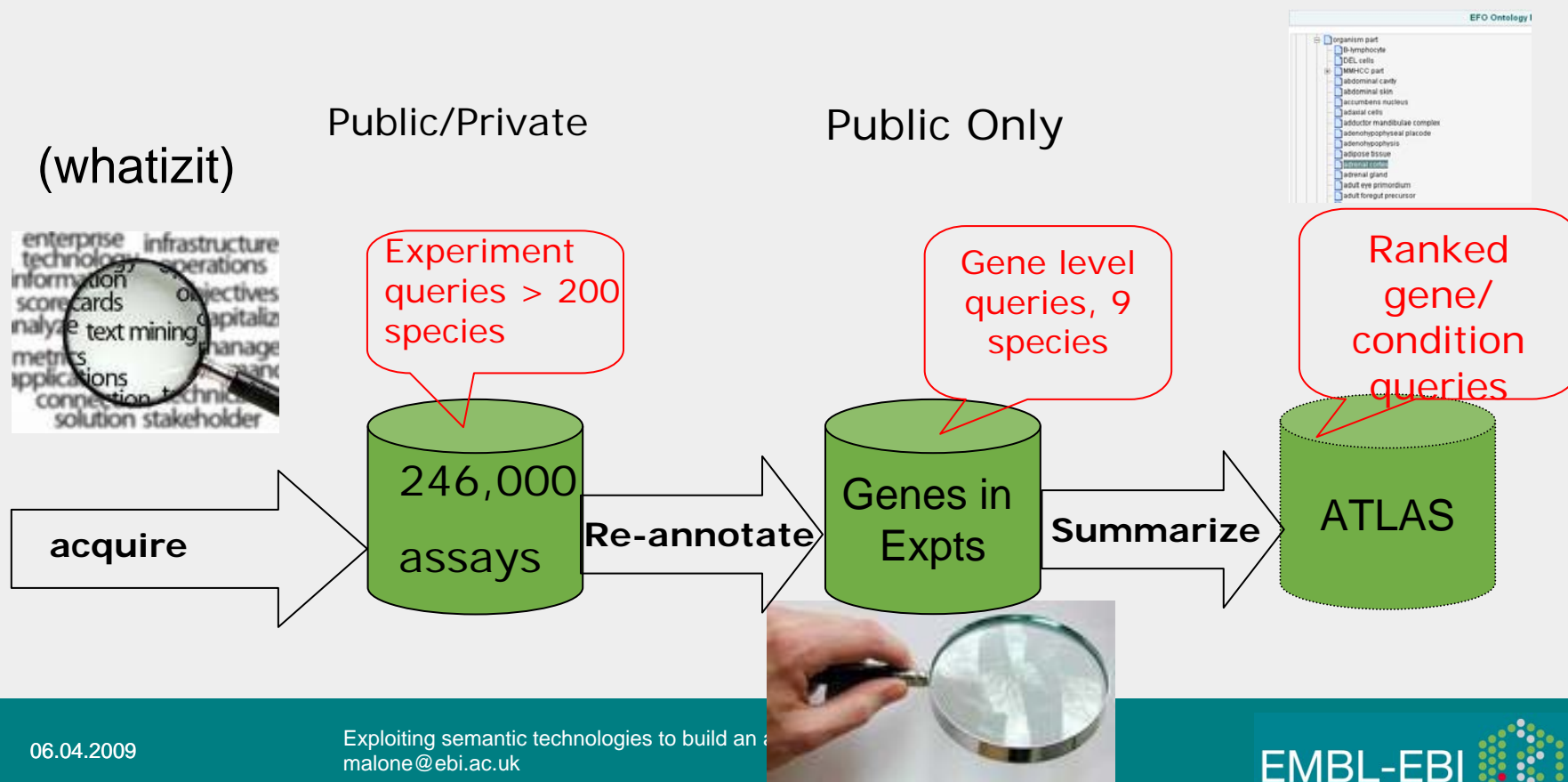
The currently selected experimental factor is highlighted. To see the expression profile for another factor just click on the factor name.



Done

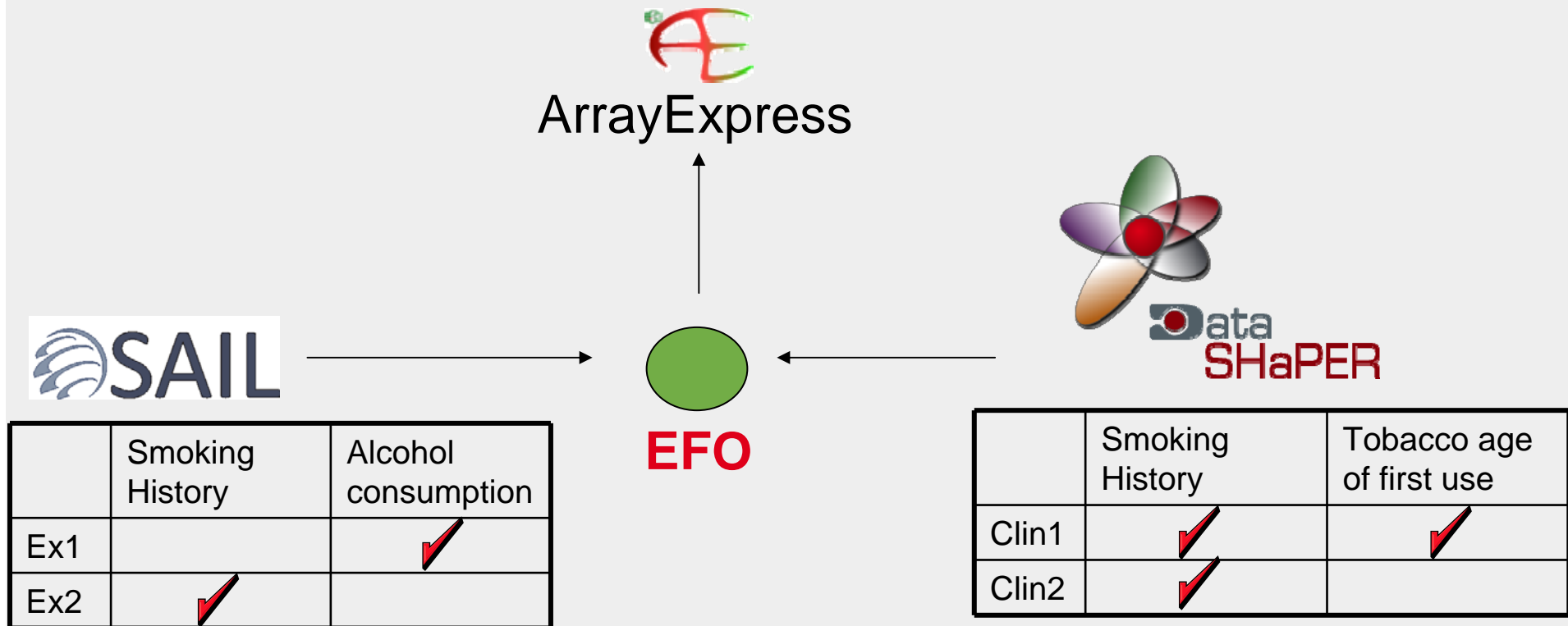
# Annotating High Throughput Data

- Text mining at data acquisition
- Ontology driven queries
- Data mining



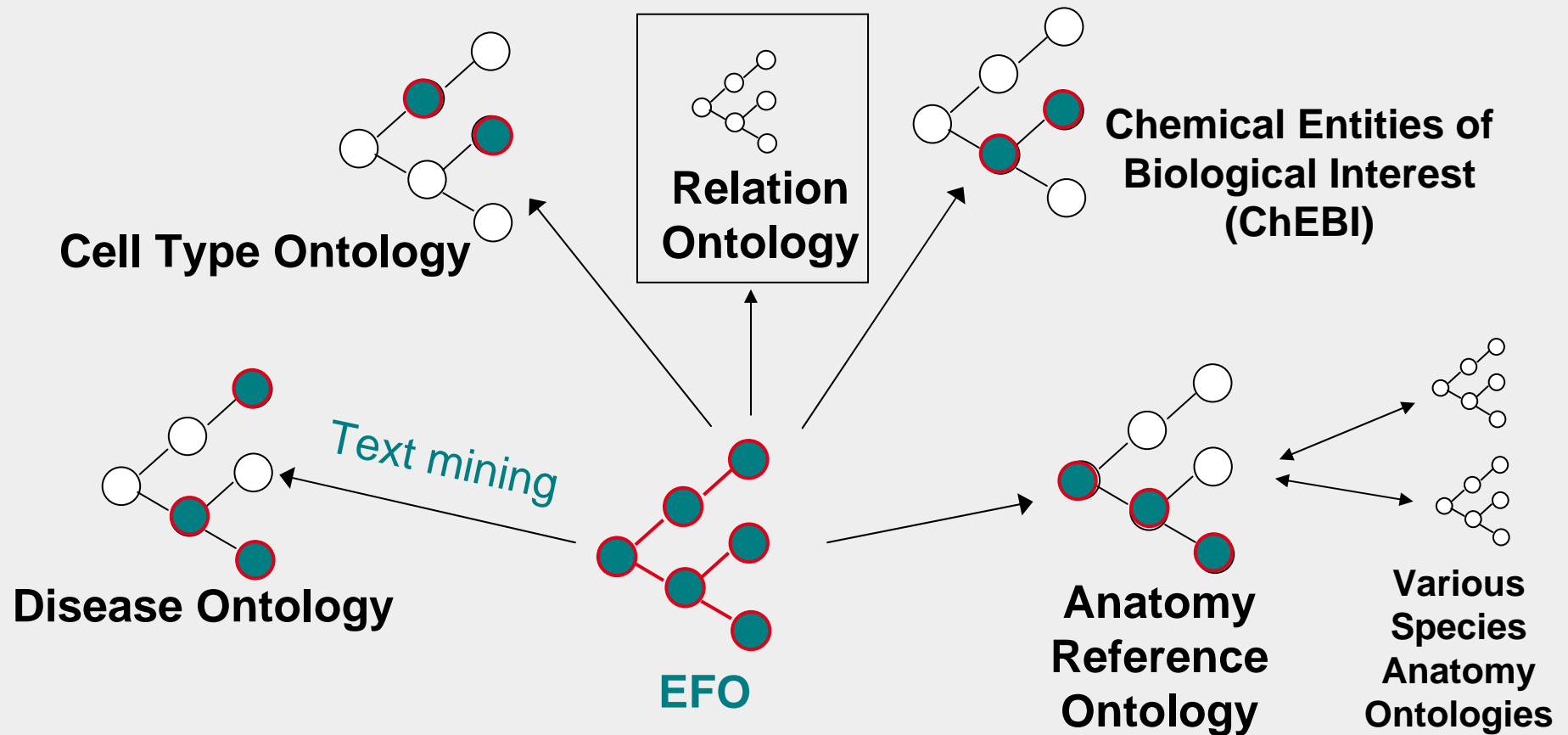
# Integrating Clinical Data

- Use cases include:
  - Primary aim: expose our gene expression experimental data
  - Secondary aim: harmonizing clinical data for study designs (e.g. GWAS)



# Building the Experimental Factor Ontology

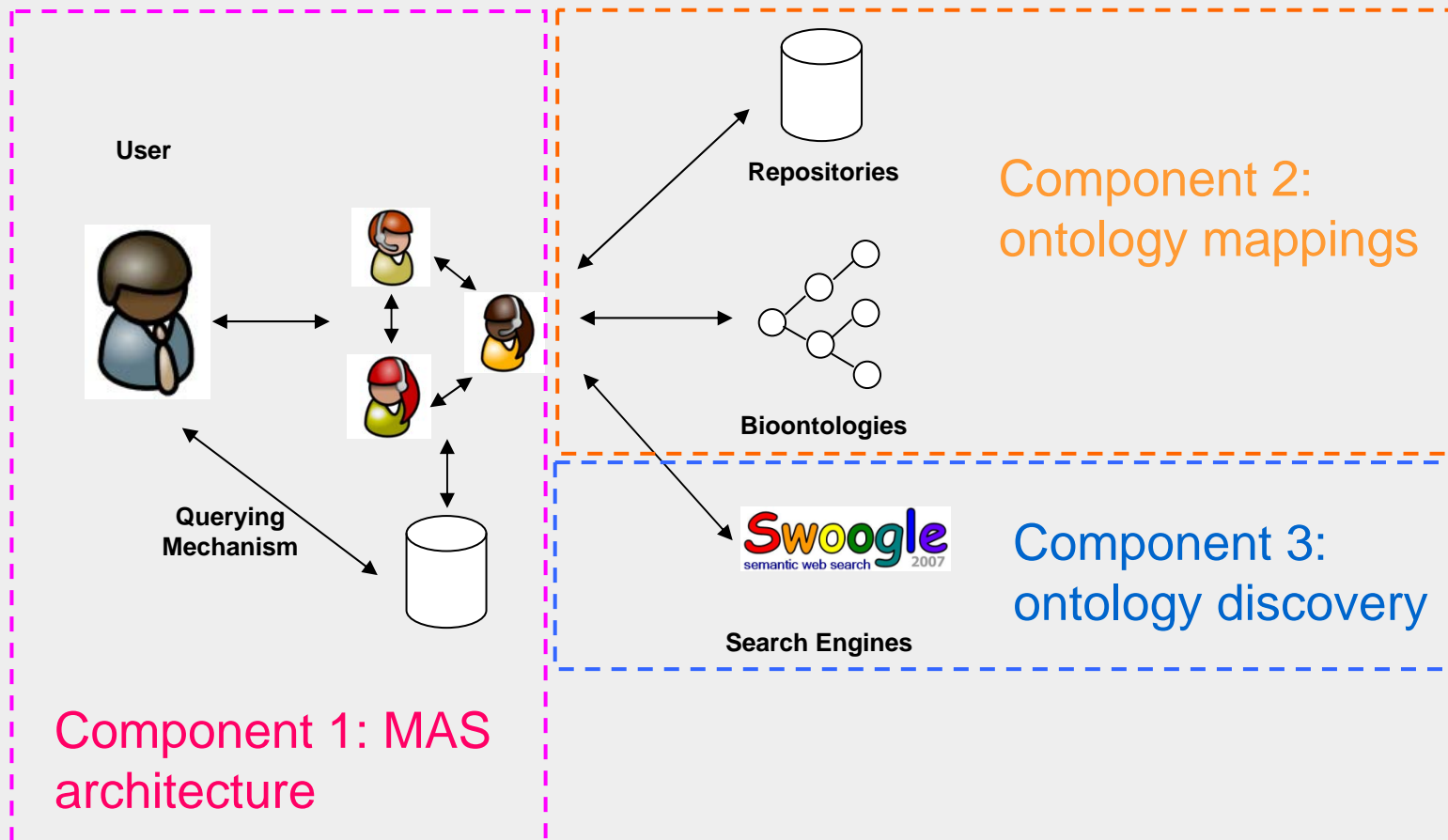
- Position of EFO in the 'bigger picture'
- Key is orthogonal coverage, reuse of existing resources and shared frameworks



# Semi-automated mapping text to ontology

- Following an evaluation from Tim Rayner we selected Double Metaphone algorithm
- Perform matching of our values in database to ontology class labels and definitions.
- Also perform mappings from EFO to other ontologies, so that EFO: cancer = NCI: cancer, DO: cancer et al.
- Sanity checking over mappings before adding to ontology

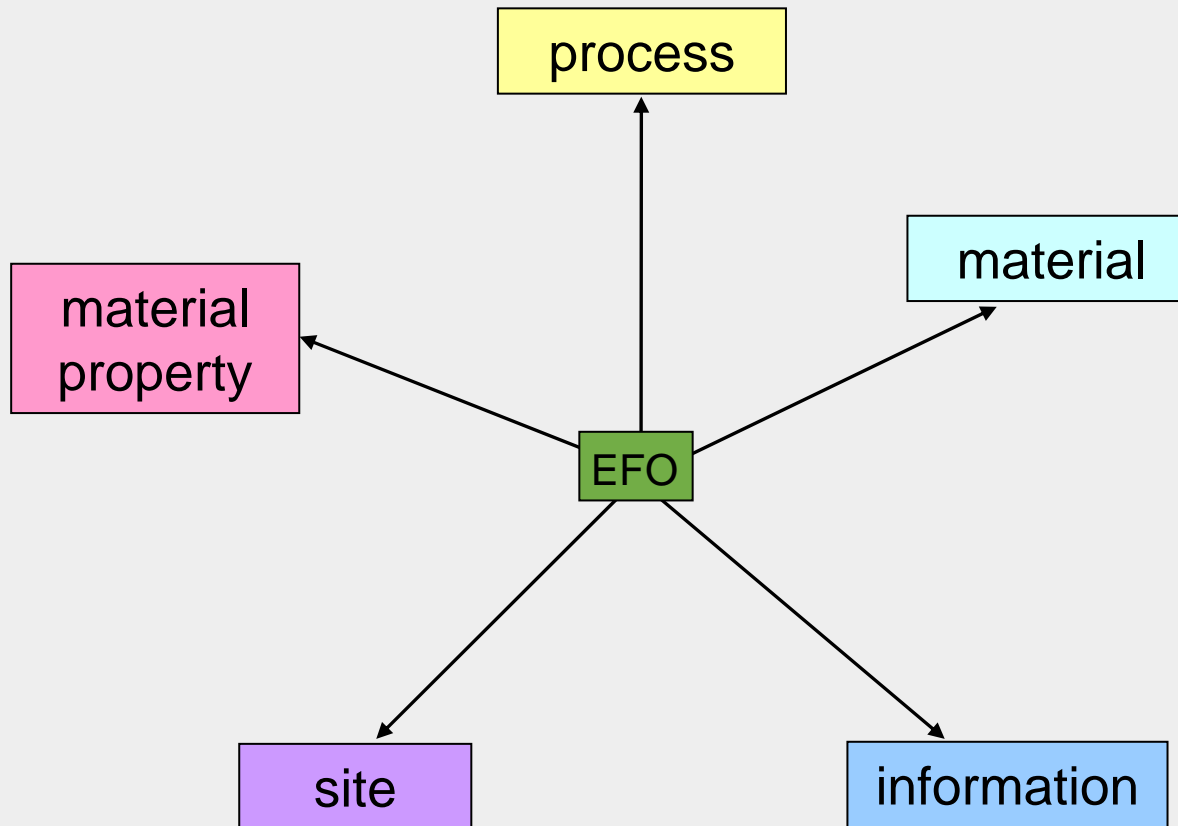
# Mapping using Agent Technology



# What does agent technology buy us?

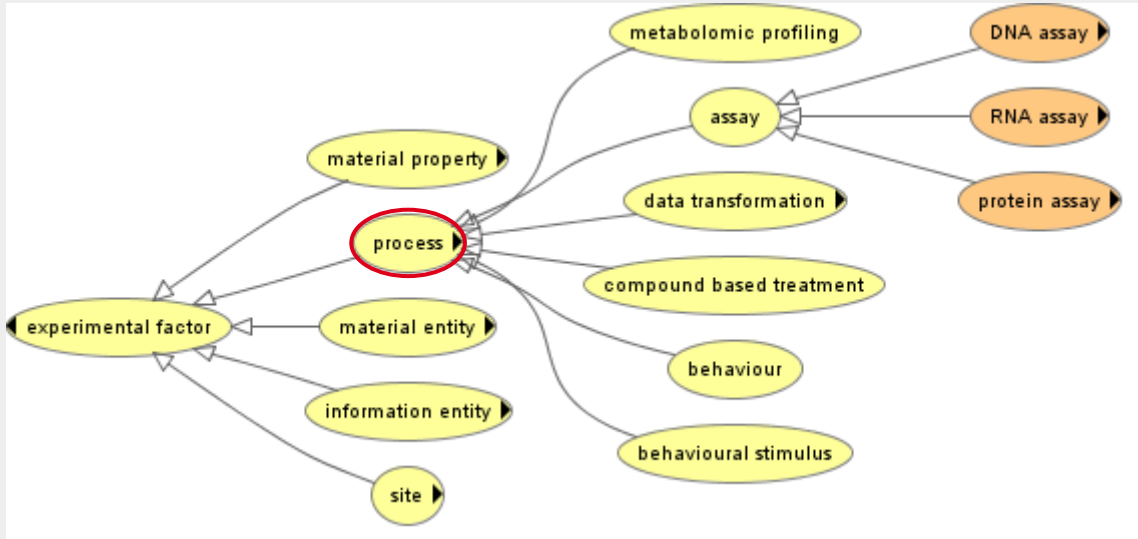
- Annotation consistency
- EFO\_1001214 is now **inconsistent**  
because DO\_15654 has new parent
- Richer mappings (hence annotations)
- EFO\_1000156 can have **new mappings**  
because new cancer class found in MIT ontology
- New potentially relevant ontologies
- New ontology found relating to **molecular + pathways**
- Semantic web compatible (i.e. can be deployed as standards compliant service)

# EFO Axes



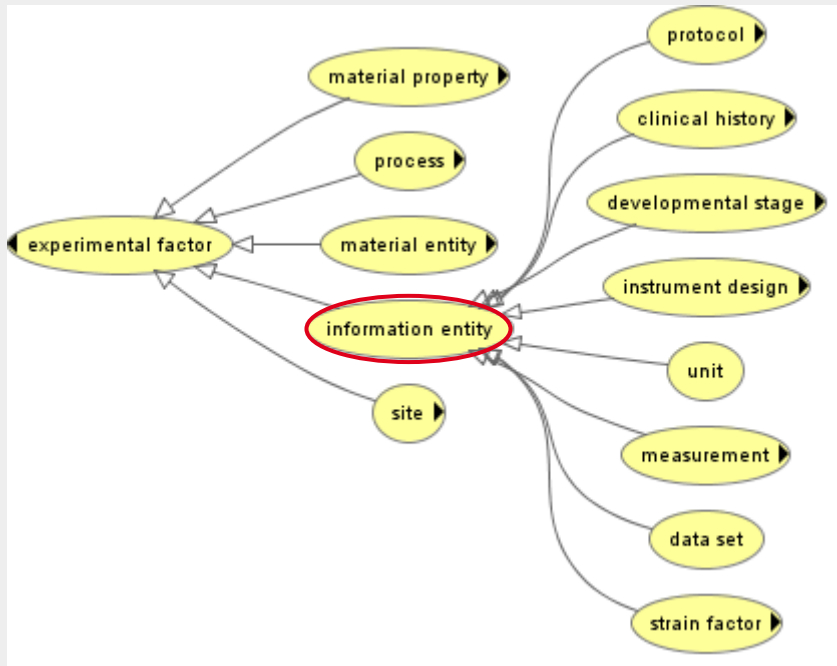
# Process

process



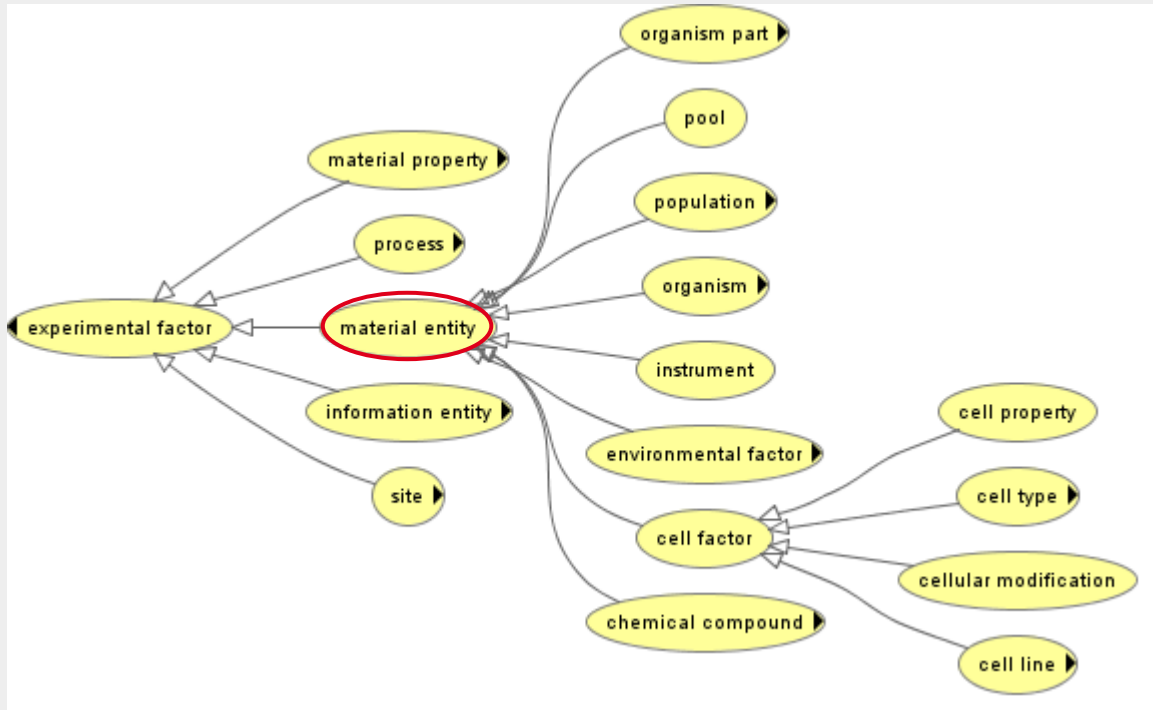
# Information

information



# Material

material



# Material Property

material property



# Using the ontology: Querying

- Public repository of gene expression data

Experiment, citation, sample and factor annotations [clear]      Filter on [reset]      Display 25

"breast cancer"      Any species       Match whole words       Loaded in ArrayExpress Atlas      Any array       Any experiment type

curator [log out]      ArrayExpress Browser Help

ID	Title	Assays	Species	Date
<input type="checkbox"/> E-GEOD-10780	Transcription profiling of human histologically normal breast tissues reveals proliferat	185	Homo sapiens	200
<input type="checkbox"/> E-GEOD-8977	Transcription profiling of human stromal samples prepared from 15 normal/DCIS anc	22	Homo sapiens	200
<input type="checkbox"/> E-GEOD-2034	Transcription profiling of human <b>breast cancer</b> samples - relapse free survival	286	Homo sapiens	200
<input type="checkbox"/> E-GEOD-10797	Transcription profiling of human breast epithelium and stroma in normal reduction ma	66	Homo sapiens	200
<input type="checkbox"/> E-GEOD-11078	Transcription profiling of human cohort of lymph node-negative <b>breast cancer</b> patient	23	Homo sapiens	200
<input type="checkbox"/> E-GEOD-11259	Transcription profiling of mouse BALB/c strain animals injected with <b>breast cancer</b> 4T:	9	Mus musculus	200
<input type="checkbox"/> E-TABM-657	Transcription profiling of human breast epithelial tissue	6	Homo sapiens	200
<input type="checkbox"/> E-MEXP-2065	Transcription profiling of human <b>breast cancer</b> cell lines treated with dasatinib and he	8	Homo sapiens	200
<input type="checkbox"/> E-GEOD-14548	Transcription profiling of human tumor microenvironment during <b>breast cancer</b> progr	66	Homo sapiens	200
<input type="checkbox"/> E-TABM-390	Transcription profiling of human dormant and angiogenic <b>breast cancer</b> , osteosarcom	8	Homo sapiens	200
<input type="checkbox"/> E-TABM-631	Transcription profiling time series of human estrogen-responsive human <b>breast cancer</b>	27	Homo sapiens	200
<input type="checkbox"/> E-GEOD-7700	shRNA profiling of human normal breast and <b>breast cancer</b> cell lines to identify genes	10	Homo sapiens	200
<input type="checkbox"/> E-GEOD-11429	Transcription profiling of human basal-like <b>breast cancer</b> - in vitro and in vivo Analy	10	Homo sapiens	200
<input type="checkbox"/> E-GEOD-12777	Transcription profiling of human (51) <b>breast cancer</b> cell lines	51	Homo sapiens	200
<input type="checkbox"/> E-GEOD-1617	Transcription profiles of mouse mammary gland development identifies estrogen res	16	Mus musculus	200

# Using the ontology: Atlas Querying

Genes:  e.g. ASPM, \"p53 binding\"  
 Organism:    
 Conditions:  e.g. liver, cancer, diabetes  [advanced search](#)

Your query was expanded via [EFO](#), an ontology of experimental variables developed by ArrayExpress Production Team

## REFINE YOUR QUERY

Genes 1-100 of 328 total found

### Experiment

[E-GEOD-3202 \(326↑ 326↓\)](#)  
[E-AFMX-5 \(324↑ 324↓\)](#)  
[E-TABM-276 \(314↑ 310↓\)](#)  
[E-AFMX-6 \(302↑ 291↓\)](#)  
[E-MEXP-561 \(295↑ 293↓\)](#)

### Organism part


[lung \(139↑ 201↓\)](#)  
[cerebellum \(146↑ 190↓\)](#)  
[tumor \(311↓\)](#)  
[caudate nucleus \(204↑ 94↓\)](#)  
[white blood cells \(297↓\)](#)

### Compound treatment

[none \(258↑ 136↓\)](#)  
[FLAP \(324↓\)](#)  
[MK886 \(307↑\)](#)  
[HNRPA2B \(292↑\)](#)  
[methotrexate \(221↓\)](#)

### Disease state

[normal \(303↑ 305↓\)](#)  
[breast carcinoma \(261↑\)](#)

Legend:  - number of studies the gene is **over/under** expressed in



Gene	Organism part									
	adrenal gland phre...	adrenal cortex	adrenal gland	pancreas	pituitary	pituitary gland	prostate	salivary gland	thymus	thyroid gland
<b>APLP1</b>	3	3	2	3	2	5	5	5	3	2
<b>TTYH1</b>	2	4	1	4	3	3	5	6	3	2
<b>CTNND2</b>	2	5	2	2	5	3	6	3	2	2
<b>SIGLEC6</b>	2	2	5	2	5	1	6	3	1	2
<b>NRXN1</b>	2	2	5	2	1	5	2	5	3	2
<b>CDH1</b>	3	6	1	1	6	4	2	3	2	2
<b>CTNNA2</b>	2	3	4	2	1	2	3	5	3	2

# Using the ontology: Exposing data via external resources

- NCBO Bioportal

## Experimental Factor Ontology

### Melanoma (Get a direct link to this concept in BioPortal)

Details	Visualization	Notes	Mappings	Resources
 Gene Expression Omnibus DataSets				A gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval. Annotations:20
 ArrayExpress				ArrayExpress is a public repository for microarray data, which is aimed at storing MIAME-compliant data in accordance with MGED recommendations. The ArrayExpress Data Warehouse stores gene-indexed expression profiles from a curated subset of experiments in the repository. Annotations:30
<a href="#">Transcription profiling of human non-invasive, invasive, AND IN-vivo passaged (very-invasive) cancer cell lines were compared FROM breast carcinoma (MCF7, MDA-MB231, 231MFP), ovarian carcinoma (OV-CAR3, SK-OV3, SK-OV3sc), AND melanoma (Mum2c, C8161, C8161sc) reveal dysregulated ether lipid metabolism drives Fra-1-dependent cancer pathogenesis</a>				
ID: E-GEOD-10709		Annotation Context: description		
<a href="#">Transcription profiling of human non-invasive, invasive, AND IN-vivo passaged (very-invasive) cancer cell lines were compared FROM breast carcinoma (MCF7, MDA-MB231, 231MFP), ovarian carcinoma (OV-CAR3, SK-OV3, SK-OV3sc), AND melanoma (Mum2c, C8161, C8161sc) reveal dysregulated ether lipid metabolism drives Fra-1-dependent cancer pathogenesis</a>				
ID: E-GEOD-10709		Annotation Context: name		
<a href="#">Adenovirus-mediated E2F-1 gene transfer in melanoma cell line</a>				
ID: E-GEOD-1562		Annotation Context: description		

# Using the ontology:

## ISACreator (www.ebi.ac.uk/bioinvindex)

ISACreator - Beta

file view help

**OVERVIEW**

- study1
  - s\_study1.txt
  - a\_micro\_trans\_profile...

**study description**

study identifier: study1

study title: a new study

study submission date: [ ]

study public release date: [ ]

study description: a test study

**study design descriptors**

+ Add new design...

Field Name	design	design	design	design
Design Type				

**study publications**

+ Add new publication...

Field Name	publication
PubMed ID	
Publication DOI	
Publication Author list	
Publication Title	
Publication Status	

**study factors**

+ Add new factor... select from previous factors

Field Name	factor	factor
Factor Name	dose	compound
Factor Type	EFO:experime...	EFO:experime...

**study assays**

+ Add new assay...

**ontologylookup**

recommended search | all ontologies

term: experimental search

- CL - Cell Type
- DOID - Human Disease
- EFO - ArrayExpress Experimental Factor
  - experimental autoimmune encephalomy
  - experimental design << 0000485 >>
  - experimental factor << 0000001 >>
- EV - eVOC (Expressed Sequence Annot
- MI - Molecular Interaction (PSI MI 2.5
- MP - Mammalian Phenotype

recent history

- UO:microliter
- PATO:male
- OBI:data transformation
- OBI:growth
- EFO:genotype
- NEWT:Saccharomyces cerevisia
- OBI:protein extraction
- OBI:auto injector;OBI:chromat
- EFO:experimental factor

selected term(s): [ ]

close ok

**STUDY - STUDY OVERVIEW**

**INFORMATION**

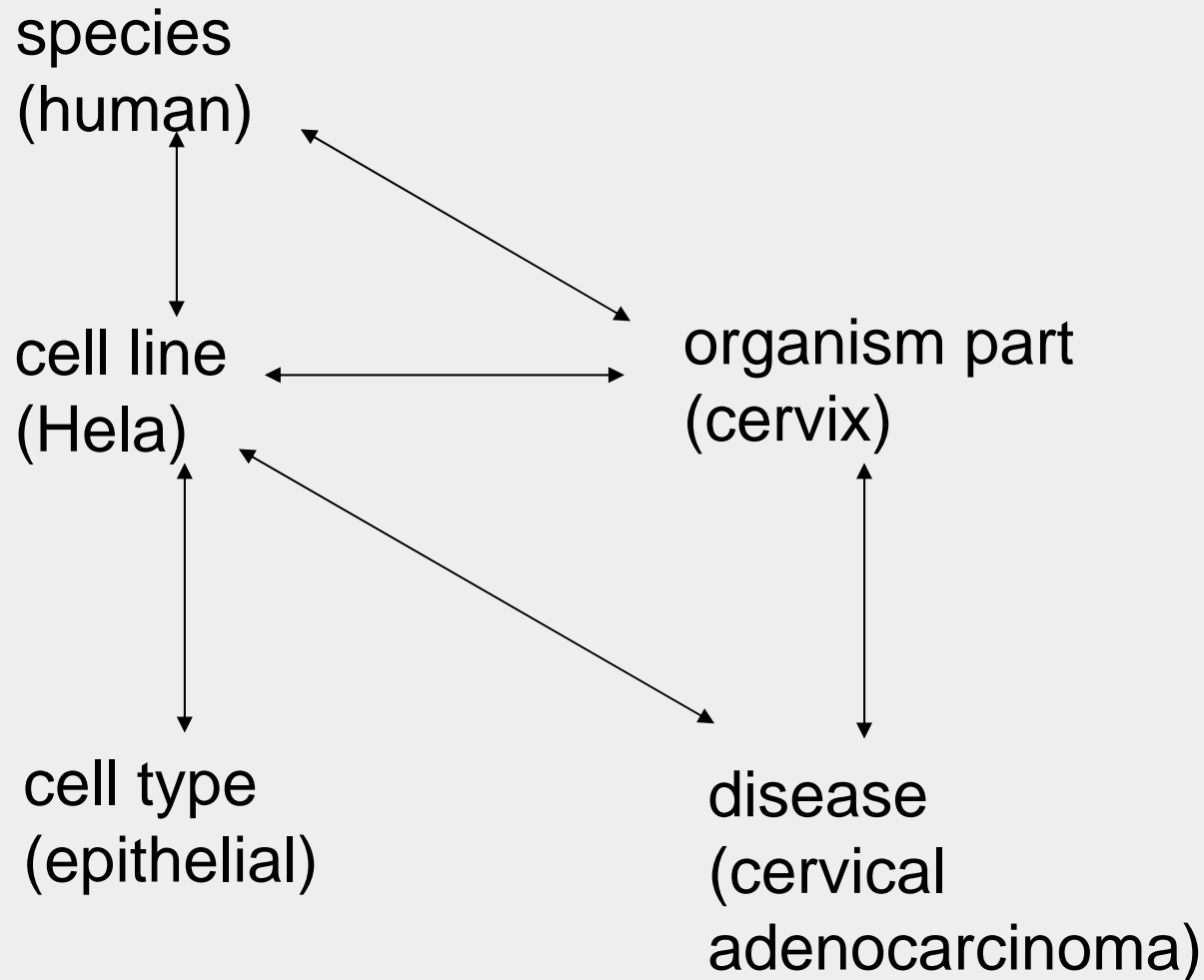
**ontologylookup**

double click on this cell to access the ontology lookup facility. use this to select controlled terms!

controlled terms provide a way to accurately describe data, experimental workflow, protocols and so forth to others...

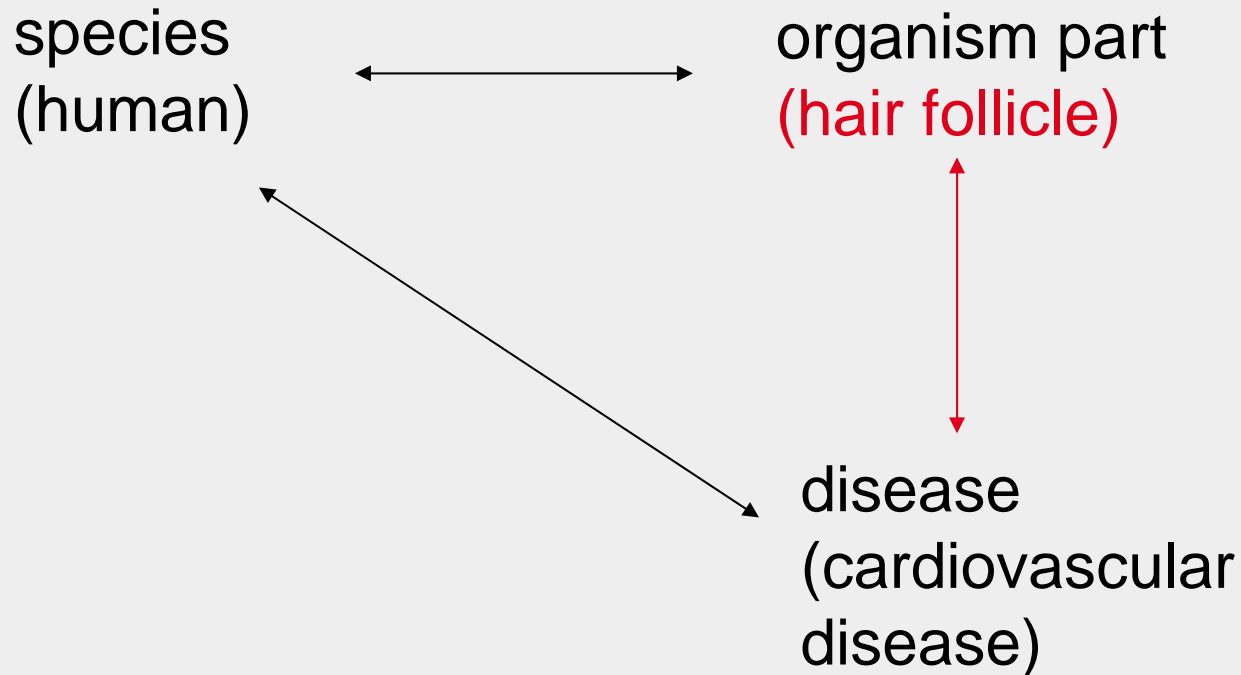
# Using the Ontology:

## Detecting Nonsense: Enforcing correctness



# Using the Ontology:

## Detecting Nonsense: Enforcing correctness



# Future Work for EFO

- Mapping in external ids– Snomed-CT, FMA, ChEBI, Brenda tissue ontology – and on requested
- API development for serving external ids from ArrayExpress Atlas API
- Working with external ontologies to produce validated cross products
- Extensions for clinical data integration Gen2Phen, Engage
- Extensions for mouse model of human disease queries
- Addressing ‘temporal dimension’
- Addition of units
- Improving query implementation in ArrayExpress Atlas – GUI changes
- Addition of synonyms
- Semantic clustering of experiments

# Conclusion

- Ontology development for text mining, annotation, query  
Built with our needs in mind, however covers a wide range of experimental variables across a wide range of technologies, extensible, open source
- Xref'd to existing ontology resources when possible
- Text mining works, reduces the workload
- 1.0 is released on April 1st 2009
- 0.10 version currently available in OLS and NCBO bioportal
  - <http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=EFO>
  - <http://www.ebi.ac.uk/microarray-srv/efo/>

# Acknowledgments

- Ontology creation:
  - **James Malone** ([malone@ebi.ac.uk](mailto:malone@ebi.ac.uk)), Helen Parkinson, Tomasz Adamusiak, Ele Holloway
- Mapping tools and text mining evaluation:
  - Tim Rayner, Holly Zheng
- External Specialist Review:
  - Trish Whetzel, Jonathan Bard
- AE Team:
  - Alvis Brazma, Anna Farne, Ele Holloway, Margus Lukk, Eleanor Williams, Tony Burdet, Misha Kapushesky
- EBI Rebholz Group (Whatizit text mining tool)
- EC (Gen2Phen, FELICS, MUGEN, EMERALD, ENGAGE, SLING), EMBL, NIH