Adventures in integrating ‘omics’ data

Helen Parkinson PhD

P3G Workshop, Luxembourg

EMBL-EBI

Overview

- Use cases – what we want to do and why

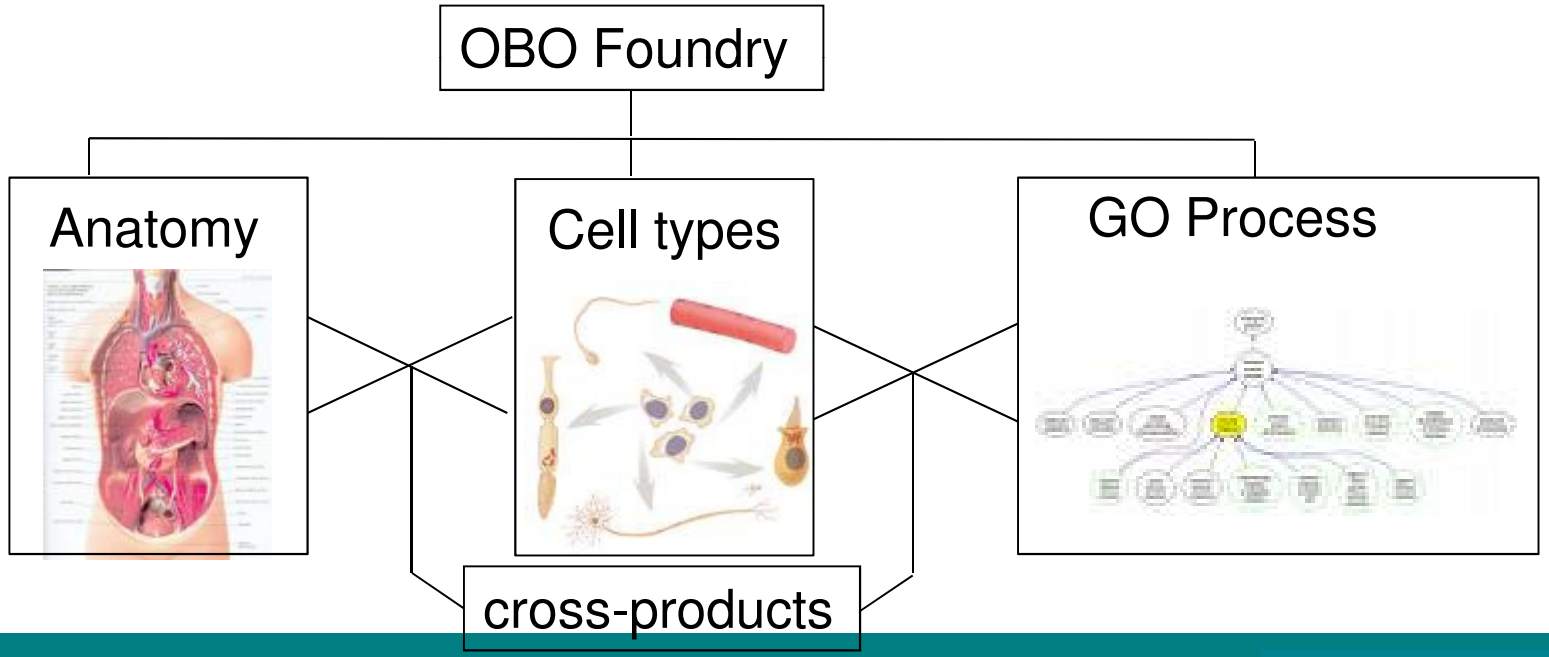
- Ontology context – what are the materials we start with
- Motivation
- Ontology and application
  - Semi-automated mapping and manual curation
  - Evaluation and application to data
  - Results
  - Desiderata for representing phenotypes

EBI data

- 1,000,000 sample annotations in ArrayExpress
- Seq DBs, tissues, metagenomics, reactions, etc
- Cross database integration issues EGA/AE/ERA etc
- Name value pairs ‘Disease’ =‘cancer’
- Algorithms, software, methods,
- Parameter annotation e.g. Virtual Physiological human
- Complex phenotypes, clinical information
- Embedded literature, pubmed abstracts, full text papers, supplemental information
- Most of the data relate to cell lines, tissues, disease samples, clinical information and phenotypes
- Millions of records, legacy data, automation needed

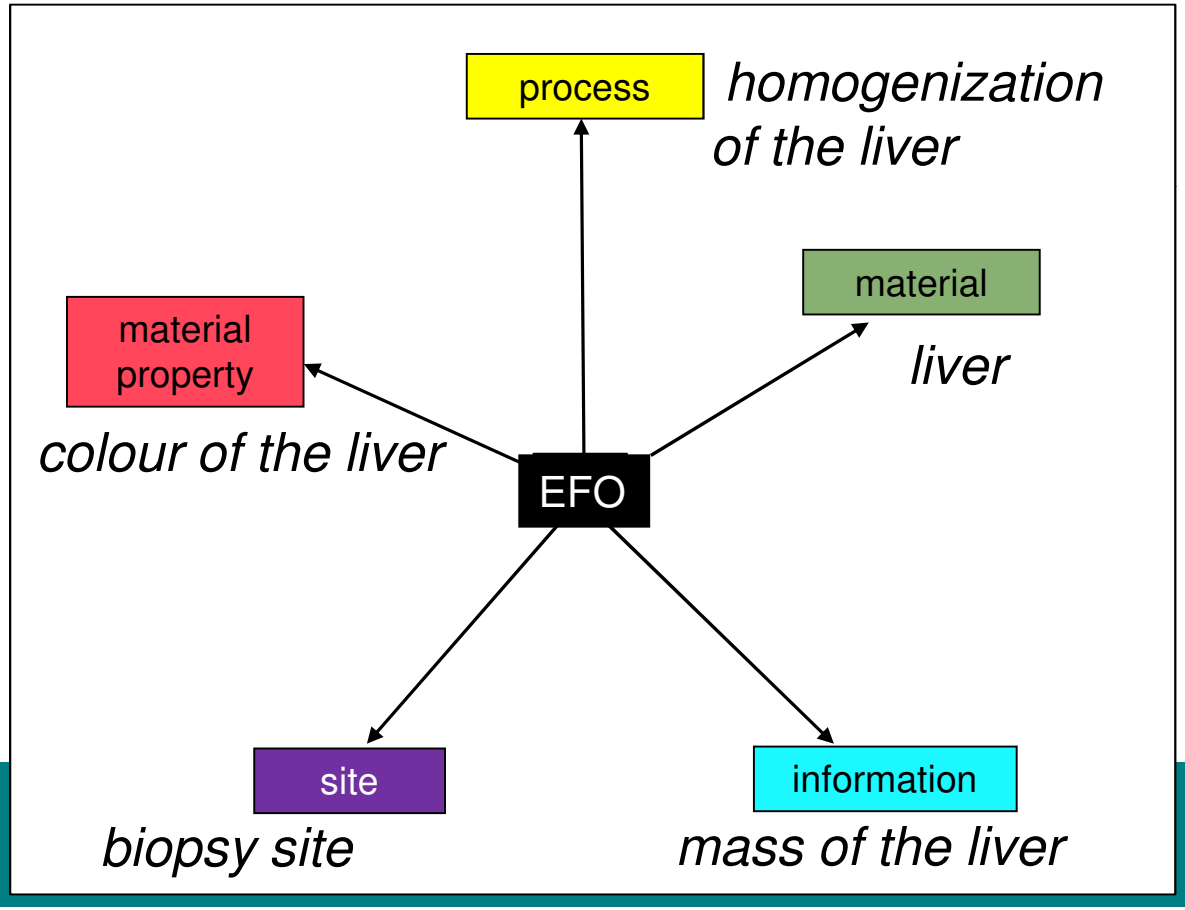
Different kinds of ontologies - Canonical

- Ontologies that represent *knowledge space*
  - Clear scope e.g. 'Normal processes'
  - And purpose – annotation of gene products
  - Applied for more e.g. Enrichment analysis and text mining
  - (Mostly) orthogonal – there is only one Cell Type Ontology
  - Foundational or Canonical Ontology



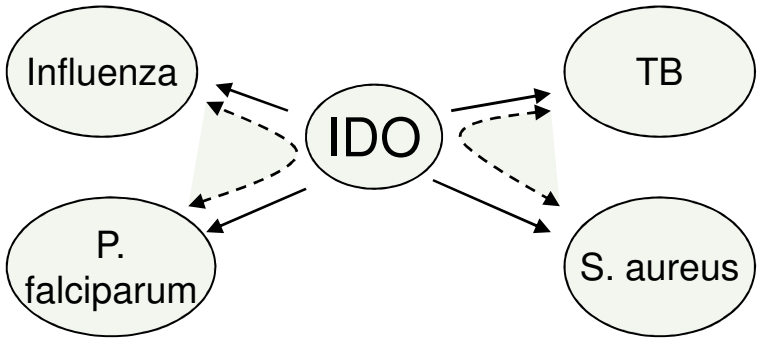
Upper level ontologies, DOLCE, NULO, BFO BioTop.....

- Much philosophical discussion - BFO in the ascendant – assumes a realist view
- ‘Ontologies mirror reality and provide domain knowledge’
- EFO is *BFO-ish*

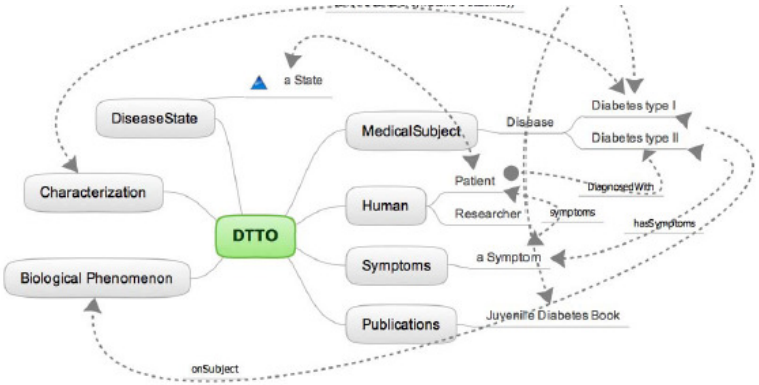


Application ontologies

- Designed to map into data e.g. EFO Type 2 Diabetes
- Typically are cross domain, not orthogonal with OBO foundry
- Consume parts of other ontologies
- Not necessarily representing reality, or knowledge - rather tools for managing, analysing and querying data
- Clear scope and range
- Testable use cases
- Typically designed with some implementation in mind



Lindsay Cowell



Eric Neumann, Pfizer

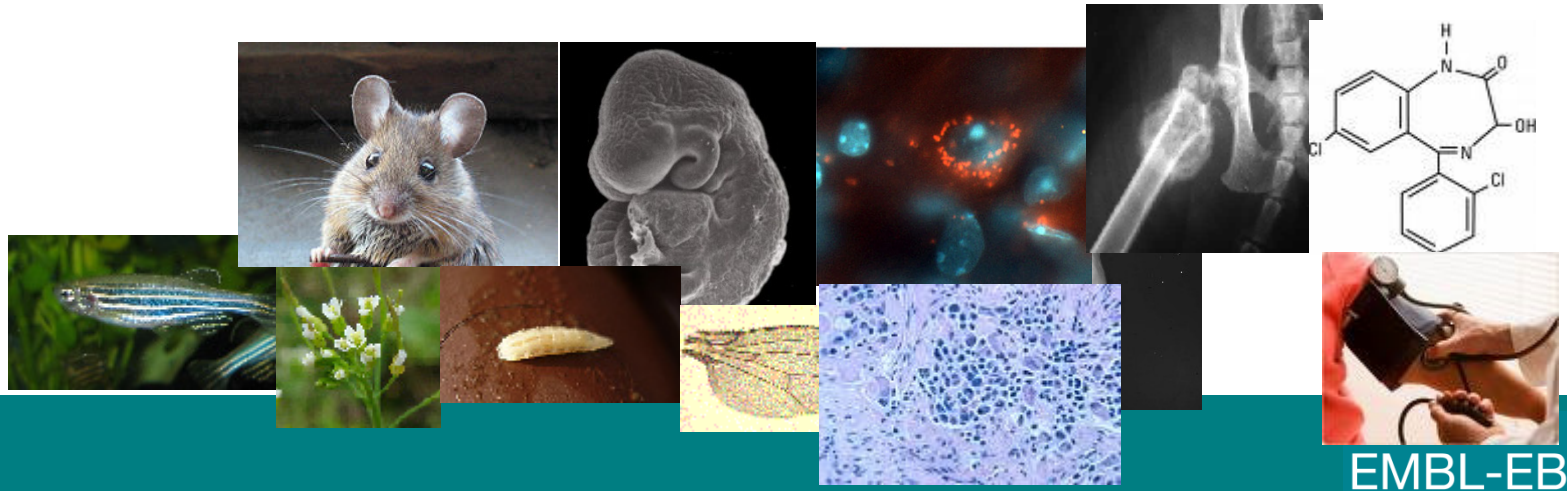
ArrayExpress Use Cases

- Query support and expansion
- Data visualization and exploration
- Summary level data presentation
- Data integration via ontology terms
- Semantic distance queries across experiments
- Cross products between – cell lines, tissues, cell types, diseases ...

- Intelligent template generation for different experiment types in submission or data presentation
- Detection of annotation inconsistency
- Annotator support, term suggestion
- Text mining at acquisition/submission for GEO data and post-hoc
- Literature text mining

Defining scope

Annotations	Archive	Atlas
Species	330	9
Samples	238,000	34,650
Annotations on samples	860,700	101830
Unique sample annotations	37,500	**6600**
Assays (Hybridizations)	246,000	30,000
Annotations on assays	569,700	67,000
Unique assay annotations	25,000	**4000**



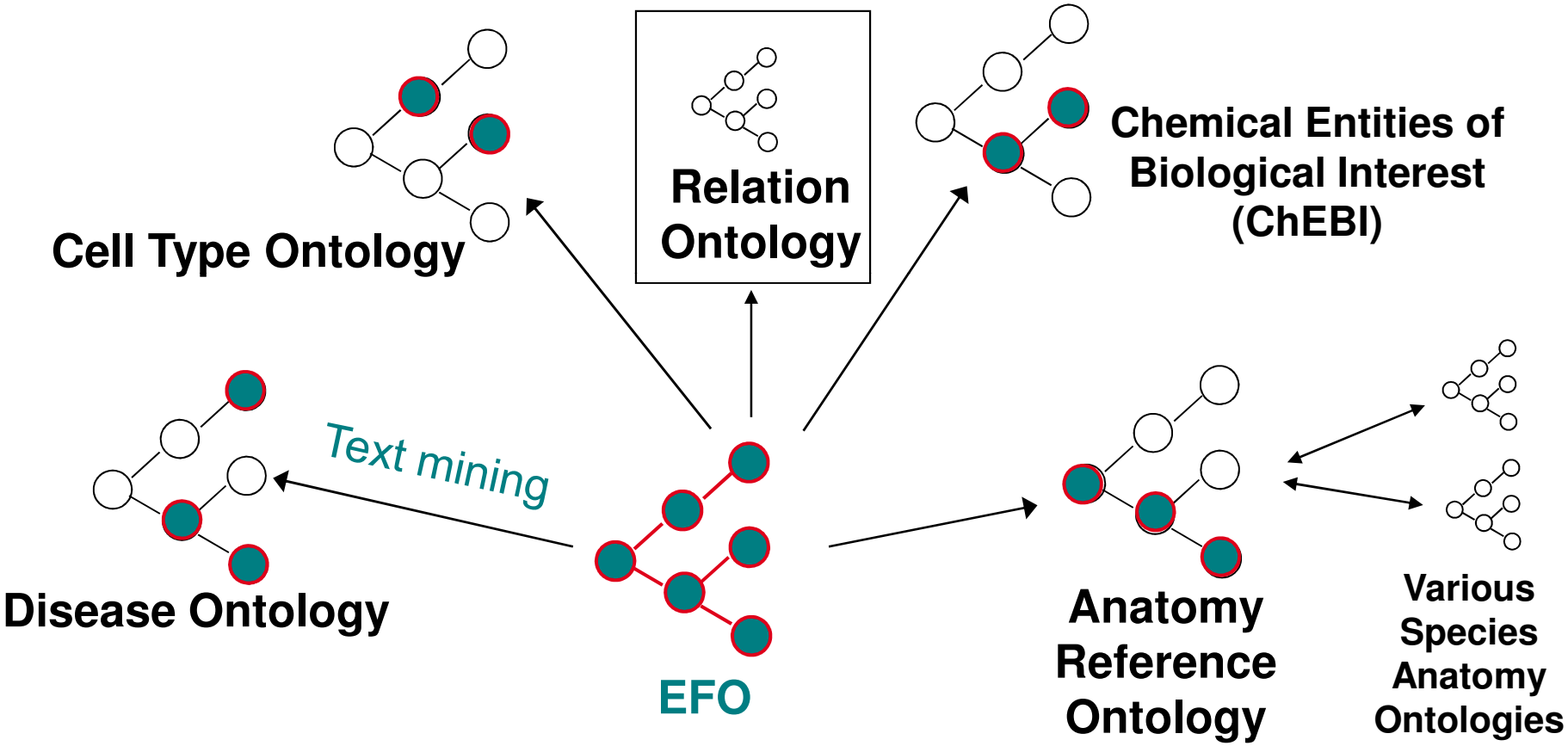
EMBL-EBI

EFO Vital Statistics

- September - version 1.4
- 14 successive monthly releases
- 2000 classes
- Built in Protégé 3.6, OWL, converted to OBO
- Available via OLS, BioPortal, www.ebi.ac.uk/EFO
- Focus on : diseases, cell types, cell lines, 'mammalian anatomy', plant terms, experimental processes, compounds,
- Mapped to:
  - Drosophila Gross Anatomy ontology, Cell Type ontology, National Cancer Institute Thesaurus,
  - Disease Ontology, Zebrafish Anatomy and Development, CRISP Thesaurus Version,
  - The Arabidopsis Information Resource (TAIR), The Jackson Lab mouse anatomy,
  - Foundational Model of Anatomy, Brenda, ChEBI, MGED ontology, Unit Ontology.

Building the Experimental Factor Ontology

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



Mapping the data and creating EFO

H E1

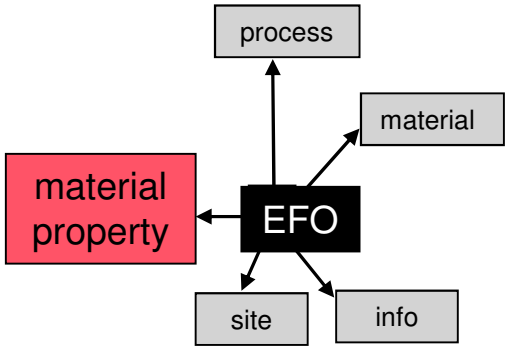
- Double metaphone algorithm for semi- automated mapping to existing ontologies
- Selected for good coverage of the data – mammalian, cancer, mouse models of disease, ...
- Annotations mapped to ontology class labels and synonyms
- EFO v0.1 created
- EFO mapped to other ontologies, so that EFO: cancer = NCI: cancer, DO: cancer *etc*
- Sanity checking mappings
- Build a hierarchy for EFO, change the backend database, insert mappings, modify the GUI
- Check and iterate, maintain
- Atlas July 2009 ~100,000 annotations
- Extend to the entire data archive – in progress

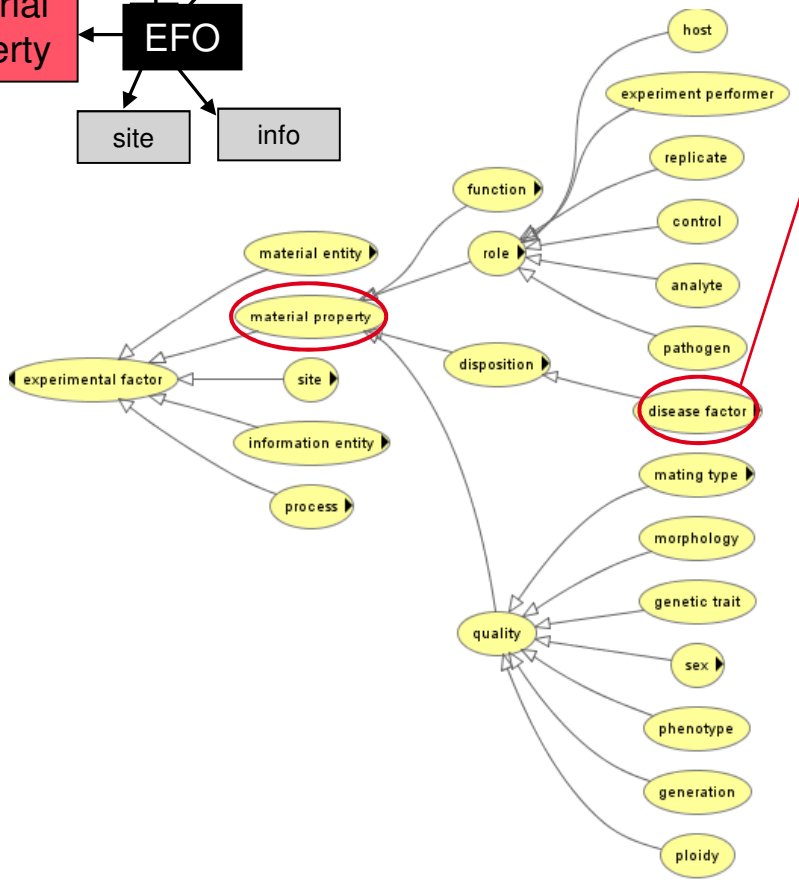
Slide 11

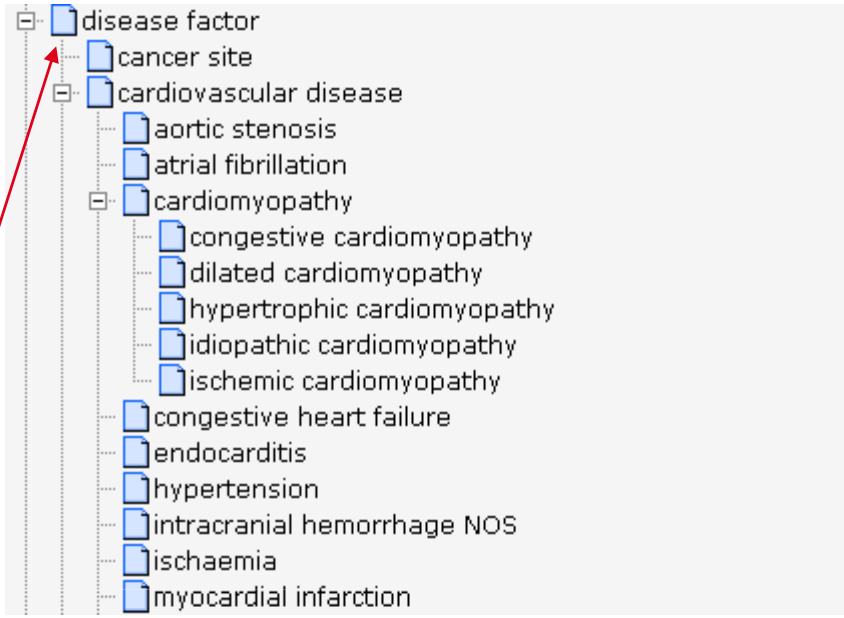
H E1 could talk a bit about the ontology being the issue not the algorithm here. Better ontology=better performance. e.g. NCIT good coverage of cancer sets, fma good for anatomy

Parkinson, 27/03/2009

Material Property







What EFO is not

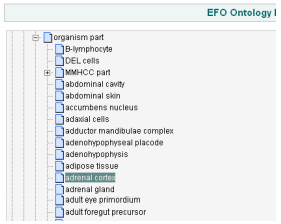
- ... not orthogonal to OBO foundry ontologies
- ... not structuring knowledge space
- ... not an automatic ontology mash-up (like Uberon)
- ....not subscribing to any particular philosophy
- ... not intended to contain numerical values
- ... Not a replacement for a generic mammalian anatomy
- ... Not using anyone else's hierarchy or classification of e.g. Disease
- ... Not a replacement for OBI, incorporates some OBI concepts and provides use cases

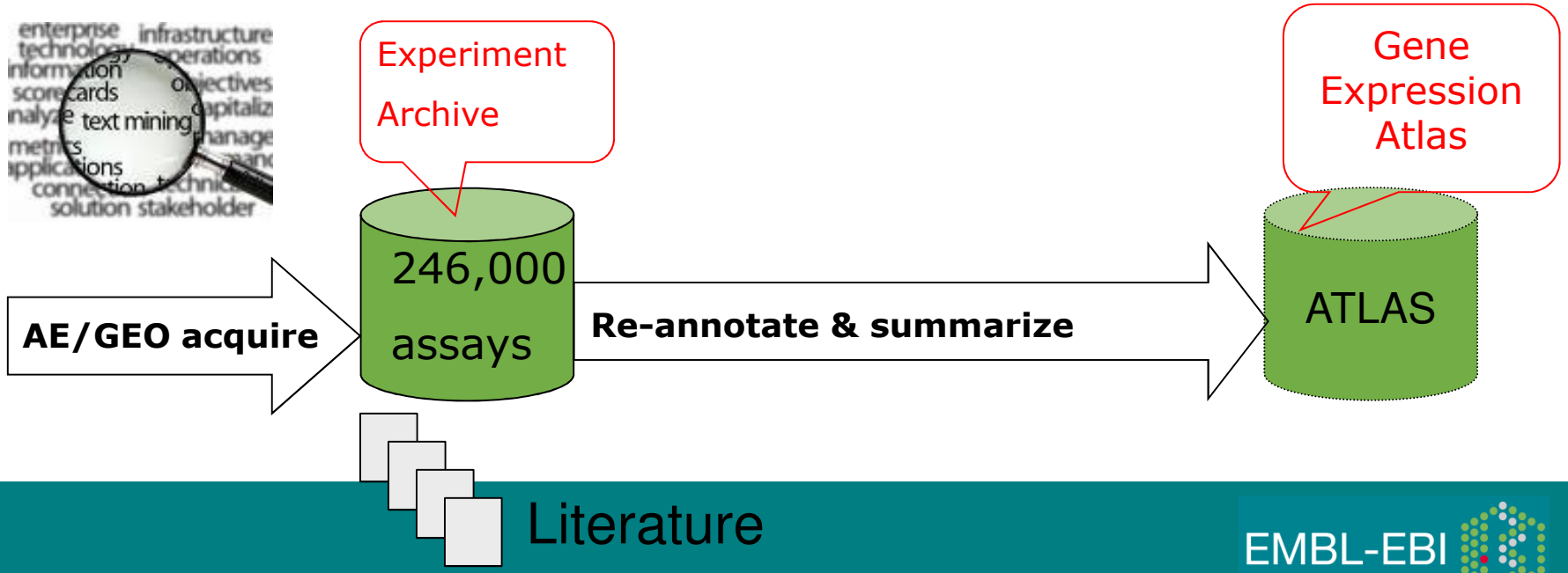
Evaluating EFO

- Does it meet our specific use cases?
- Can we deploy it in a GUI and annotation tools?
- Does it make sense for our users?
- Can we easily maintain and extend it?
- Can other people use it?
- Can we develop it further for new applications?
- Is it ontologically robust?

Annotating High Throughput Data

- Text mining at data acquisition ✓
- Ontology driven queries ✓
- Data mining ✓
- Data driven ontology development ✓
- Text mining other people's data – in progress
- Text mining literature – in progress





Clean data - Atlas Querying

e.g. Cell adhesion genes in all ‘organism parts’

‘View on EFO’

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

Gene	Organism	experimental factor	sample factor	organism part	animal component	animal developme...	extraembryonic tis...	placenta	umbilical cord	animal reproductiv...	female reproductiv...	animal ovary	female reproductive system	appendage	vertebrate limb	digit	cardiovascular sys...	heart	cardiac atrium	myocardium	ventricle	craniofacial tissue	eye	lacrimal gland	retina	mouth	tongue	nose	digestive system c...	intestine	large intestine	colon
Kitl	Mus musculus	53/55	41/39	29/28	29/28					4/3	3/1	2	1		1	1/2	1/2				3	3		2	2	1	1		1				3/2	3/2	2	2
Myh9	Mus musculus	54/50	40/35	29/26	29/26	1	1	1	1	2/4	2/1	1	1		1	3	3				3/1	3/1	1	1	1	1/1	1		1			1	2/1	2/1	2	2
Cx3cl1	Mus musculus	39/44	34/33	27/27	27/27					3	1	1				3	3				2/1	2/1	1	1	2	1/1	1/1	1	1				6/2	6/2	1/2	1/2
Scarb1	Mus musculus	54/50	39/39	28/26	28/26					4/3	2/2	1/1	1		1	2/1	2/1				2/1	2/1		1	1/1	1/1	1/1	1	1				1/5	1/5	2	2
Cd34	Mus musculus	47/46	35/34	27/26	27/26	1	1		1	4/3	3/2	2/1	1	1	1	1/1	1/1	1	1	1	3/1	3/1	1	1	2	1/2	2	1	1	1	1	1	2/3	1/3	1	1
Mpdz	Mus musculus	49/52	35/41	25/27	25/27					2/4	2/1	1	1		1	3	3				1/2	1/2		1	2	2	2		2				4	4	2	2
Vcam1	Mus musculus	57/54	42/38	27/25	27/25					3/2	2	2				3	3				2/1	2/1	1	1	1	1/1	1/1	1	1				3	3	2	2
Ezr	Mus musculus	53/48	41/31	27/24	27/24					2/3	2/1	1	1		1	2	2				3/1	3/1	1	2	2	2	2	1	1				4	4	2	2
Cd24a	Mus musculus	53/56	42/39	25/26	25/26					2/3	2	2		1		3	3				2/4	2/4	1	2	2	2	2	1	1	1	1		5/2	5/2	4	3
Ssx2ip	Mus musculus	48/46	37/37	25/24	25/24	1	1		1	4/2	2	1		1		4	4				1/3	1/3		2	2	1/1	1/1	1	1				3	3	2	2

EMBL-EBI

Dirty data

Experiment, citation, sample and factor annotations [clear]

cel

- cell line
- cell type
  - cardiac myocyte
  - platelet
  - mononuclear cell
  - neuronal stem cell
  - skin cell
  - erythroblast
  - mesangial phagocyte
  - interneuron
  - fibroblast
  - parietal cell
  - mesenchymal stem cell
  - male germ cell
  - glial cell (sensu vertebrata)

Filter on [reset]

Any species

Any array

Any experiment type

Display options

25 exper

Detailed vie

		Assays	Species	Date
	cells cultured in the presence of HA-pul	8	Mus musculus	2009-09-15
	sitive and negative human B lymphocyt	14	Homo sapiens	2009-09-12
	asts (CAFs) reveals are activation in in	6	Mus musculus	2009-09-12
	with and without TGF-B vs human annu	13	Homo sapiens	2009-09-12
	vascular endothelial growth factor, ant	4	Homo sapiens	2009-09-12
		36	Homo sapiens	2009-09-12
	om patients with OA undergoing total k	12	Homo sapiens	2009-09-12
E-GEOD-14554	Transscription profiling of rat primary hepatocytes toxicogenomic comparison of TCDD and PCB	26	Rattus norvegicus	2009-09-12
E-GEOD-14553	Transcription profiling of primary human hepatocytes - toxicogenomic comparison of TCDD and	0		2009-09-12
E-GEOD-14340	Transcription profiling of human neural crest cells	5	Homo sapiens	2009-09-12
E-GEOD-12843	Transcription profiling of human mesenchymal stem cells derived from adult adipose and lipoma	4	Homo sapiens	2009-09-12
E-GEOD-12806	Transcription profiling of human Dendritic cells reveals inhibition of Chlamydia pneumoniae repli	4	Homo sapiens	2009-09-12
E-GEOD-11686	Transcription profiling of human wrist muscles from cerebral palsy patients	16	Homo sapiens	2009-09-12
E-GEOD-17812	Transcription profiling of mouse memory P14 T cells with control or mutated ThPOK	4	Mus musculus	2009-09-11
E-GEOD-17513	Transcription profiling of mouse embryonic stem cell derived cardiac progenitors	12	Mus musculus	2009-09-11
E-GEOD-16836	Transcription profiling of human CD16+ and CD16- peripheral blood monocytes from healthy inc	8	Homo sapiens	2009-09-11

1724 experiments, 54831 assays. Displaying experiments 1 to 25. Pages: 1 2 3 4 5 6 7 8 9 10 .. 69

Java Implementation

- Atlas, terms mapped by curator and added to EFO and database, autocomplete method added
- Archive – parses the OWL file (OWL api) on the fly and lucene search-> synonym lookup ontology-expansion
  - People don't annotate their data well, tools incentives – need templates
  - 1,000,000 sample annotations as name value pairs
  - 10,000 experimental records
  - 250,000 assays
  - Free text, typos, user edited data, html markup, ...
  - Lucene indexing is robust and fast (needed a patch for phrase handling)
  - Query expansion is fast (EFO is small)
  - 7 weeks development time to date, x 1 undergraduate student, x 0.25 Senior Java programmer
  - No learning curve for our users

Views on ontologies

- OBO foundry provides canonical ontologies
- Some ontologies provide subsets – e.g. GO slims, FMA
- Tricky things about views
  - You need a well defined use case
  - Reasoning over a view
  - They are cross ontology e.g. Cells in tissues, diseases and anatomy
  - So we need views x views
  - Technology is bleeding edge
- EFO is a view, defining a view is as hard as building an ontology
- Each use case takes longer to refine than the view

Solving ~~th~~ our ~~ld's~~ ontology problems?

- Ontology development for our use cases: text mining, annotation, query
- Covers a wide range of experimental variables across a technologies, extensible, open source, inexpensive
- Xref'd to existing ontology resources where possible
- Deployed in the AE production environment
- Added value above other ontologies, new cross products
- QC and feedback for external ontologies
- Use cases for views
- Leverages OBO foundry efforts
- Monthly release cycle
- www.ebi.ac.uk/efo

Why not import terms and preserve their names spaces (MIREOT) a la OBI

1. MIREOT what? Too much choice
   Cell from FMA, cell type ontology, GO?

2. We add axioms, is it the same class once we've added new parents, or annotation properties?

3. We often have no axioms at source, but when we do, what do we do when we do have axioms, hard to recode these

Future work

- Software ontology
- Mapping EFO into OBI @1.0
- Cross study semantic similarity queries – prototyped
- API to allow query access using your ontology id, not ours and visualisation in context
- Template mark up – generic representations of common cases in tools (Annotare)
- Support for the Phenotype model (Morris's ppt)
- Applying these techniques to other data sets

Conclusions and desiderata

- Good public domain semantic resources
  - Some ontologies well formed, some being fixed
  - Actively maintained, useable
  - We lack cross ontology mappings and we need experts for that
- Many groups building their own application ontologies
- Need better technologies for generating views
- Mapping between ontologies, cross products, - all need use cases
  - Main barrier is lack of human/mouse anatomy/phenotype mapping
- We shouldn't need to do these every time for every case
- Building new ontologies is easy and not (always) desirable

- Model mapping
- Test data sets, marked up with one or more ontologies e.g. protocols and OBI
- Use cases, mappings, added value ontologies
- Test ontologies on data, define preferred ontologies for this community, build if needed
- Easy(ier) access to resources like OBI – useage, pros and cons – manual
- Improve the collection of new data
- Mine the legacy data

EFO Acknowledgments

- **Ontology creation:**
  - **James Malone, Tomasz Adamusiak, Ele Holloway, Helen Parkinson, *Jie Zheng (U Penn) = 1FTE***
- **Ontology Mapping tools and text mining evaluation:**
  - ***Tim Rayner*, Holly Zheng, Margus Lukk**
- **GUI Development**
  - **Misha Kapushesky, Pasha Kurnosov, Anna Zhukova. Nikolay Kolesinkov**
- **External Review and anatomy:**
  - **Jonathan Bard, Jie Zheng**
- ArrayExpress Production Staff
- EBI Rebholz Group (Whatizit text mining tool)
- Many source ontologies for terms and definitions esp. Disease Ontology, Cell Type Ontology, FMA, NCIT, OBI
- Funders: EC (Gen2Phen,FELICS, MUGEN, EMERALD, ENGAGE, SLING), EMBL, NIH
- Eric Neumann, Joanne Luciano and Alan Ruttenberg
- OBI developers