The Experimental Factor Ontology Utilities in Knowledge Discovery and Sharing

Sirarat Sarntivijai, Tony Burdett, Simon Jupp, and Helen Parkinson

http://www.ebi.ac.uk/efo/

siiraa@ebi.ac.uk

@SSirarat
Outline

• Ontologies at the EBI
• Using and building ontologies with the real data at the Centre for Therapeutic Target Validation (CTTV)
• The ontology toolkit
• Challenges
Data resources at EMBL-EBI

**Genes, genomes & variation**
- European Nucleotide Archive
- 1000 Genomes
- Ensembl Genomes
- European Genome-phenome Archive
- Metagenomics portal

**Gene, protein & metabolite expression**
- ArrayExpress
- Metabolights
- Expression Atlas
- PRIDE

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

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

**Chemical biology**
- ChEMBL
- ChEBI

**Systems**
- BioModels
- Enzyme Portal
- BioSamples

**Literature & ontologies**
- Europe PubMed Central
- Gene Ontology
- Experimental Factor Ontology
- BioSamples Database

**Cross domain resources**
- IntAct
- Reactome
- MetaboLights

**+~10, BAO, CTO, ORDO, ChEBI, Fly Anatomy, SWO, OBI,…**

**Reactions, interactions & pathways**
The Experimental Factor Ontology

- We build the EFO in OWL
- EFO is an **application ontology**, built for use in production services
- Lots of external dependencies - reuses where possible and appropriate other references
EFO reuses existing knowledge
Working the abundance of data

- **BioMedBridges**, the pilot project
  - A LOT of data, ALL TYPES of data
- **Need shared standards and semantic web technologies**
- **CORBEL** scaling up BioMedBridges
  - **COordinated** Research infrastructures
  - **Building** **Enduring** **Life-science services**
CTTV Core: Target Validation Platform

The Core components provide the backbone of information and display for the CTTV platform. The key challenge is to integrate data sources that affect the validity of a target in a single infrastructure, allowing seamless interrogation of these data. This integration involves both the more conceptually easy (though often practically far harder) integration of related information, and also the more complex joint computation of information together.

CTTV Computational Pipelines

The Pipelines component is the information being fed into the CTTV core database and web interface. Much of the utility and flexibility of the Core will be enabled by ensuring that data are described in a machine readable form and are queryable in a standard way that allows access to pertinent, particularly clinical, information including phenotypic relationships and distinctions. To enable this we will use ontologies (core vocabularies) as infrastructure to provide sets of common terms and defined synonyms, and to specify relationships between terms (e.g.
I am interested in target T:  I am interested in disease D:

Which diseases can be treated by modulating target T?  Which targets can be modulated to treat disease D?

What's new on release 0.4?

See all changes in the release notes document.
Major update of ChEMBL disease associations
Improved associations for GWAS, Expression Atlas and EVA
New filters for data types and pathways in the display of targets associated with a disease
New visualization for variants, gene isoforms and genomic context

FAQ

See all the frequently asked questions
What can I search for?
Where does the data come from?
How is the association score calculated?
EFO in CTTV data analysis

Diagram showing relationships between diseases such as asthma, non-small cell lung cancer, squamous cell carcinoma, osteosarcoma, Crohn's disease, ulcerative colitis, atopic eczema, seasonal allergic rhinitis, osteosarcoma, ulcerative colitis, celiac disease, and serum ST2 measurement.
CTTV Semantic interoperability

- We can now ask questions of meta data attached to
  - Variants, known drugs, proteins, genes, pathways
  - We can link common to rare disease via shared phenotypes
  - We have used Uberon, GO, ChEBI, ORDO, OBI, etc in the context of EFO to do this
  - It scales, and it gives good coverage of the sources
  - Next: regulatory information and COSMIC
Lessons learned and solutions
The reality of data - Lots of stuff, lots of types of stuff

<table>
<thead>
<tr>
<th>Annotations</th>
<th>Gene Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td>~1821</td>
</tr>
<tr>
<td>Samples</td>
<td>1,126,457</td>
</tr>
<tr>
<td>Annotations on samples</td>
<td>7,672,825</td>
</tr>
<tr>
<td>Unique sample annotations</td>
<td>223,650</td>
</tr>
<tr>
<td>Assays (Hybridizations)</td>
<td>965,638</td>
</tr>
<tr>
<td>Annotations on assays</td>
<td>3,248,298</td>
</tr>
<tr>
<td>Unique assay annotations</td>
<td>189,381</td>
</tr>
</tbody>
</table>

This is top 200. There are nearly 250,000 in total and half are used only once

gender:female

sex:female

disease:breast cancer

frequency=2285
Data comes in many forms

- We need fuzziness which is hard to do programmatically

<table>
<thead>
<tr>
<th>Input from submission</th>
<th>Ontology class</th>
</tr>
</thead>
<tbody>
<tr>
<td>2’-deoxy-5-azacytidine</td>
<td>5-aza-2’-deoxycytidine</td>
</tr>
<tr>
<td>Ovarian Cancer</td>
<td>ovarian carcinoma</td>
</tr>
<tr>
<td>Anterior tibialis</td>
<td>tibialis anterios</td>
</tr>
<tr>
<td>Barret’s Esophagus</td>
<td>Barrett’s esophagus</td>
</tr>
<tr>
<td>Endothelium, Vascula</td>
<td>cardiovascular system endothelium</td>
</tr>
</tbody>
</table>
Solutions – the toolkit
Ontology is often used as a backbone infrastructure for data mapping.

- Experimental data are unstructured.
- Map them once, the structure sustains.
- Tools for knowledge discovery and transfer are developed with ontologies.
Data – Map them once, the structure sustains.

Multiple Views

External ontologies

Ontology development

protégé

Webulous

Analysis/Visualisation

Database curators

Data submission
Tools for knowledge discovery and transfer are developed with ontologies.
Zooma – exploiting our curators

• Our data is manually curated – (e.g. Expression Atlas, GWAS Catalog, Reactome)
  The data is aligned to ontology classes by experts

• We would like to exploit the years of work

• Zooma aims to reuse previous annotations to inform future annotations

• We assume curators are correct (though we can fix things we think may be incorrect, sometimes we find ontology errors)

• Seeded with data from Atlas, ArrayExpress and the NHGRI GWAS catalog and the ontology - EFO
Zooma annotation model – simple example
<table>
<thead>
<tr>
<th><strong>long cell phenotype</strong></th>
<th><strong>Automatic</strong></th>
<th><strong>CMPO_0000077</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>lamellipodia + high actin ratio cells</strong></td>
<td><strong>more lamellipodia cells, increased number of actin filament phenotype</strong></td>
<td><strong>Automatic</strong></td>
</tr>
<tr>
<td><strong>lamellipodia cells</strong></td>
<td><strong>more lamellipodia cells</strong></td>
<td><strong>Automatic</strong></td>
</tr>
<tr>
<td><strong>sm phenotype</strong></td>
<td><strong>metaphase arrested phenotype, round cell phenotype</strong></td>
<td><strong>Automatic</strong></td>
</tr>
<tr>
<td><strong>reduction in ir-induced 53bp1</strong></td>
<td><strong>site of double-strand break</strong></td>
<td><strong>Automatic</strong></td>
</tr>
<tr>
<td><strong>cell death</strong></td>
<td><strong>cell death phenotype</strong></td>
<td><strong>Automatic</strong></td>
</tr>
<tr>
<td><strong>cell migration</strong></td>
<td><strong>cell migration phenotype</strong></td>
<td><strong>Automatic</strong></td>
</tr>
<tr>
<td><strong>Dynamic changes</strong></td>
<td><strong>increased variability of nuclear shape in population</strong></td>
<td><strong>Automatic</strong></td>
</tr>
<tr>
<td><strong>failure in decondensation</strong></td>
<td><strong>absence of mitotic chromosome decondensation phenotype</strong></td>
<td><strong>Requires curation</strong></td>
</tr>
<tr>
<td><strong>Grape</strong></td>
<td><strong>graped micronucleus phenotype</strong></td>
<td><strong>Automatic</strong></td>
</tr>
<tr>
<td><strong>cell shape processes or spiky or stretchy</strong></td>
<td><strong>regulation of nrf-kappaa transcription factor activity phenotype</strong></td>
<td><strong>Automatic</strong></td>
</tr>
<tr>
<td><strong>increased number of actin puncta or dots</strong></td>
<td><strong>increased amount of punctate actin foci phenotype</strong></td>
<td><strong>Automatic</strong></td>
</tr>
<tr>
<td><strong>increased number of zigzag actin stress fibers</strong></td>
<td><strong>increased amount of zig-zag stress fibers</strong></td>
<td><strong>Automatic</strong></td>
</tr>
</tbody>
</table>
Improving the Experimental Factor Ontology

- Dealing with term submissions
- A lot of e-mail and spreadsheets get passed around
Webulous server

- [Link](http://www.ebi.ac.uk/efo/webulous)
- Take a design pattern
- Define server side templates for cell lines
Connect to Webulous server to load templates
Sidebar for search BioPortal and creating data validations from ontology terms
Users input

- Users fill in templates and submit back to server from within Google Sheet
EFO Cell lines – OPPL patterns

- Spreadsheets converted into OWL axioms
- Validated by EFO curators and imported into live version
- UriGen server assign accessions and notify submitter
http://www.ebi.ac.uk/ontology-lookup/

You can also specify which ontology you want to search the term from.
Visualization of the hierarchy classification
A common genetically determined, chronic, inflammatory skin disease characterized by rounded erythematous, dry, scaling patches. The lesions have a predilection for nails, scalp, genitalia, extensor surfaces, and the lumbosacral region. Accelerated epidermopoiesis is considered to be the fundamental pathologic feature in psoriasis.

**Synonyms:** PITYRIASIS NEC & NOS, PUSTULOSIS OF PALMS SOLES, PUSTULAR PSORIASIS OF PALMS SOLES, Psoriasis and similar disorders (disorder), OTHER PSORIASIS, Pustulosis of Palms and Soles, Psoriasis and similar disorders, Palmoplantar Psoriasis, Pustulosis Palmaris et Plantaris, PSORIASIS RELATED DIS NEC, Other psoriasis and similar disorders excluding psoriatic arthropathy, Pustular Psoriasis of Palms and Soles, Psoriasis and similar disorders (navigational concept), Psoriases, Other psoriasis and similar disorders (disorder), Other psoriasis and similar disorders, Psoriasis and similar disorders NOS, Psoriasis and similar disorders NOS (disorder), Other and unspecified pityriasis

---

**Class relations**

**Parents**
- immune system disease
- skin disease

**Subclass of:**
- has disease location some skin
OLS3 Java script widgets

```html
<ols-autocomplete data-ontologyName="EFO"></ols-autocomplete>
```
Toolkit Summary

• Data – ontology mapping tool Zooma
• Ontology development tool – Webulous
• OLS3 – new ontology lookup service
• Portable widgets,
• APIs,
• Open code base, Apache 2.0
Acknowledgements

• Samples, Phenotypes and Ontologies Team EBI, esp. James Malone and Helen Parkinson
• CTTV Core Team, esp. Samiul Hasan, Gautier Koscielny
• GWAS Catalog Team, NHGRI, EBI, NCBI
• IMPC, EBI, WTSI, MRC Harwell
• Funders
  • NHGRI Grant Number 1U41HG007823-01
  • NHGRI, the NHLBI, and the NIH Common Fund under grant U54-HG004028.
  • EC – BioMedBridges – FP7 Capacities Specific Programme, grant agreement number 284209
  • Diachron –European Commission FP7 grant agreement 601043
  • NIH Common Fund - Grant number: U54-HG004028, EMBL-EBI, MRC Harwell, and WTSI
  • CTTV – GSK, EMBL-EBI and WTSI
  • EMBL core funds