

# Developing an application focused experimental factor ontology: embracing the OBO Community

James Malone\*, Tim F. Rayner, Xiangqun Zheng Bradley and Helen Parkinson

EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

---

## ABSTRACT

Motivation: The recent crop of bio-medical standards has promoted the use of ontologies for describing data and for use in database applications. The standards compliant ArrayExpress database contains data from >200 species and >110,000 samples used in genotyping, gene expression and other functional genomics experiments. We considered two possible approaches in employing ontologies in ArrayExpress: select as many ontologies as cover the species, technology and sample diversity, choosing where there are non-orthogonal resources and attempt to make them interoperable; or build an extensible interoperable application ontology. Here we describe the development of an application focused Experimental Factor Ontology and describe its use at ArrayExpress.

[www.ebi.ac.uk/ontology-lookup/browse.do?ontName=EFO](http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=EFO)

## 1 INTRODUCTION

The value of having explicit and rich semantic representations of data is becoming increasingly clear in bioinformatics. This is apparent in the emergence of the OBO foundry (Smith *et al.*, 2007) and numerous metadata standards (<http://www.mibbi.sf.net>). The OBO foundry promotes the development of orthogonal ontologies that are expressed in a common shared syntax, use unique namespace identifiers and explicit textual definitions for all ontology terms. These ontologies give us the terminology to describe the level of detail that content standards such as MIAME require. Underpinning this increased focus on the use of ontologies is that richer and explicit representations enhance interoperability and facilitate machine readability. As the numbers of ontologies and standards increase, the complexity of supporting standards using ontologies also increases.

In this paper we describe development of the Experimental Factor Ontology (EFO), an application focused ontology. EFO models the experimental variables (e.g. disease state, anatomy) based on an analysis of such variables used in the ArrayExpress database. The ontology has been developed to increase the richness of the

annotations that are currently made in the ArrayExpress repository, to promote consistent annotation, to facilitate automatic annotation and to integrate external data. The methodology employed in the development of EFO involves construction of mappings to multiple existing domain specific ontologies, such as the Disease Ontology (Dyck and Chisholm, 2003) and Cell Type Ontology (Bard *et al.*, 2005). This is achieved using a combination of automated and manual curation steps and the use of a phonetic matching algorithm. This mapping strategy allows us to support the needs of various ArrayExpress user groups who preferentially use different ontologies, to validate existing ontologies for coverage of real world high throughput data in public repositories and to provide feedback to the developers of existing ontologies. An additional reason to have a local application ontology – rather than simply create an enormous cross product ontology (i.e. classes created by combining multiple classes from other ontologies) – is that the structure of such an ontology may be challenging for many users and time consuming to produce (Bard and Rhee, 2004). Instead, data acquisition tools can employ one ontology rather than many external ontologies.

Brinkley *et al.* (2006) highlight the potential value in reference ontologies for performing mapping and integration for building application ontologies. However, at present these frameworks and all necessary reference ontologies do not exist. We therefore exploit the use of the several OBO Foundry ontologies as reference ontologies in contrast to the definition discussed by Brinkley *et al.* by employing a softer and more cautious view of these ontologies. Specifically, we aim to map to the concept names and definitions provided by external ontologies without importing covering axioms, thereby reducing the potential for conflict and removing an obstacle for interoperability. Instead we use references in the same way many OBO Foundry ontologies reference external resources using a pointer to their identifier. This strategy avoids ‘bedroom ontology development’ wherein ontologies are developed *ab initio* without considering the reuse of existing ontologies. By re-using and mapping we leverage the user supplied annotations and existing ontologies.

The EFO is represented in the web ontology language (OWL) thereby conforming to an accepted common representation and we also implement a policy of unique

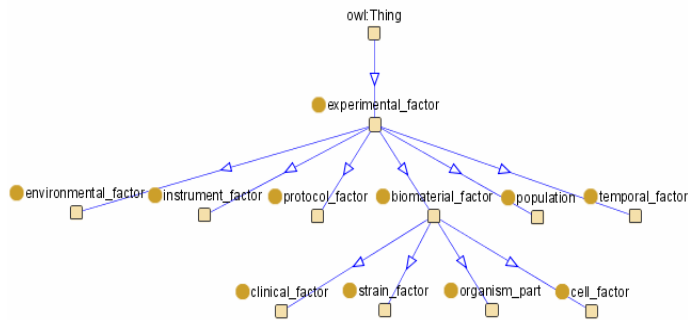
---

\*To whom correspondence should be addressed.  
Email: [malone@ebi.ac.uk](mailto:malone@ebi.ac.uk)

namespace identifiers and definitions for all terms, as encouraged by OBO. Finally, we assess our ontology *post-hoc* using semi-automated methods to assess the coverage we have obtained in terms of our set of use cases (described in our web resource <http://www.ebi.ac.uk/microarray-srv/efo/index.html>) and, hence, assess the suitability of the ontology for the task at hand.

## 2 METHODOLOGY

Since the EFO is an application ontology, we developed a well defined set of requirements based on our needs for annotating experimental data. ArrayExpress typically has ~ five annotations per biological sample, and the most important annotations are those that contain information on the experimental variables. These are both biological i.e. properties of the experimental samples (e.g. sex or anatomy), and procedural; properties of protocols used to treat the samples (e.g. sampling time or treatment with compound). The initial focus in developing the EFO is on the former as they are more likely to be present in a reference ontology (i.e. non-numeric) and can be automatically discovered in unstructured data. This is an important use case for ArrayExpress as thousands of experiments are imported from the Gene Expression Omnibus where the sample annotation is essentially uncurated free text. Additionally from analysis of user queries, biological information is more commonly queried than procedural information.



**Figure 1** High level classes from the EFO

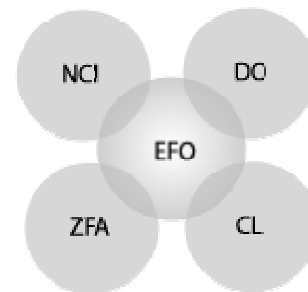
### 2.1 Mapping, curating and integrating

Our approach is a middle-out ontology methodology as described by Uschold and Gruninger (1996). In this method, we start with a core of basic terms identified from our use cases and specialize and generalize as required. Our set of initial core terms already provided some structure as the more specific concepts (called factor values) were grouped into factor categories. We then created generalized classes to give some additional structure to our ontology (shown in Figure 1). The structure at the highest level has been designed to be simple, and intuitive to biologists and

the curators, who will be the primary users of the EFO in the short term, by constructing this as an abstraction of the existing structure in ArrayExpress.

EFO terms have no internal text definitions by design, instead we leverage the mapping strategy defined below to create links to text definitions created by domain experts.

The mapping strategy involves selecting likely reference ontologies and evaluating their coverage of terms present in ArrayExpress. This includes ontologies such as the Disease Ontology which also has many mapped terms, the Cell Type Ontology and Zebrafish Anatomy and Development ontology (Sprague *et al.*, 2006) and the NCI thesaurus (Fragoso *et al.*, 2004) which has human, mouse and rat terms related to cancer (Figure 2).



**Figure 2** The intersection of the EFO and reference ontologies

To perform our mapping and add terms to EFO we used the following iterative methodology:

- Identify OBO Foundry ontologies relevant to an EFO category based on annotation use cases
- Create subset of classes of relevance to the ontology, e.g. classes under disease for disease ontology
- Perform mapping using text mining phonetic matching algorithm. This produces a list of candidate ontology class matches.
- Manually validate matched ontology classes and curate where necessary
- Manually map high quality annotations (identified as present in the ArrayExpress data warehouse) to multiple source ontologies
- Consider number of instances of terms used in ArrayExpress to determine depth and breadth
- Integrate into EFO, adding appropriate annotation values to definition and external ontology ID
- Structure EFO to provide an intuitive hierarchy with user friendly labels

### 2.2 Phonetic matching

Our matching approach uses the Metaphone (Phillips, 1990) and Double Metaphone algorithms (Phillips 2000) which were selected following an empirical study of commonly used matching algorithms and their utility in the biomedical

domain. We were particularly interested in algorithms yielding low false positive rates, as we wished to use the same algorithm for automatic annotation of incoming data.

We matched the user supplied cell type terms deposited in ArrayExpress with the Cell Type Ontology using Soundex (<http://en.wikipedia.org/wiki/Soundex>), Levenshtein edit distance (Levenshtein, 1966), Metaphone (Phillips, 1999) and Double Metaphone (Phillips, 2000) algorithms. Synonyms and term names were used during the matching process and matches were either single or multiple. For the purposes of automated annotation, single matches are obviously more desirable. The Metaphone algorithm yielded the lowest false positive rate, with 98% of the matches mapping to single ontology terms, and of these only 6% were deemed to be invalid following inspection by an expert curator. However, the overall coverage of the input term list was relatively low (17% of all terms matched). In comparison, the Double Metaphone algorithm provided higher list coverage (50% of terms) at the expense of generating a smaller proportion of single matches (48% of total matches) and a much higher false positive rate (34% of single matches). The Levenshtein and Soundex algorithms yielded results similar to the Metaphone and Double Metaphone algorithms, respectively, but both generated slightly higher levels of false positives. A combined strategy was therefore implemented, using Metaphone for a first pass and then falling back to Double Metaphone for those terms not matched by Metaphone. Using this strategy with curator supervision to select the correct term in the multiple-match cases yielded the highest overall number of matches with minimal human intervention. Verified matched terms identified by this strategy were included in the EFO and placed manually in the hierarchy.

### 2.3 Ontology conventions

Naming conventions described by Schober *et al.* (2007) were used. Specifically, class labels are intended to be meaningful to human readers, short and self-explanatory. They are singular and conform to the conventional linguistic and common usage of the term, for example, the term *Huntingdon's disease* has a capital H since it is a proper noun, whereas *cancer* would not. Identifiers have the format EFO:00000001, where a unique integer identifies a term and EFO identifies the ontology. We use an alternative term annotation property to capture synonyms for class labels, text definitions are not provided at present. The ontology is developed in Protégé and converted to OBO format for display in OLS.

## 3 THE EFO

Part of the hierarchy visualized in OLS is shown in Figure 3. The current version of EFO has ~800 child terms of the class experimental factor. The majority of these have been mapped to external reference ontologies and knowledge

resources, as indicated by the definition citation annotation property.

As an early version, the ontology still has parts that are under review and is evolving. In particular, the hierarchy still contains classes that are likely to be moved and changed to add more structure as it is relatively flat at present. Furthermore, the additional group of use case covering cross species queries, e.g. disease and mouse model of disease, and the representation of anatomical parts in different species are required but are currently not supported by the EFO. However, as the iterative engineering process is ongoing, these will be addressed in the near future. Where possible we will use existing resources to address these use cases.

### 3.1 Validation

The ArrayExpress data flow doubles on a yearly basis. This allows us to constantly validate the ontology against fast changing annotation with a variety of granularities. It also allows us to develop the ontology against emerging use cases. We have implemented an iterative evaluation of the ontology against the data content of the ArrayExpress repository, against newly submitted data for curation purposes and also against the ArrayExpress data warehouse – a set of additionally annotated and curated data which represents the ArrayExpress ‘gold standard’. As the ontology evolves it will be used daily by the ArrayExpress production team and incremental versions will be tested internally prior to public release. Early stage evaluation is performed semi-automatically by mapping between the ontology and very large meta-analyzed curated experiments and by comparison with reference ontologies. We were able

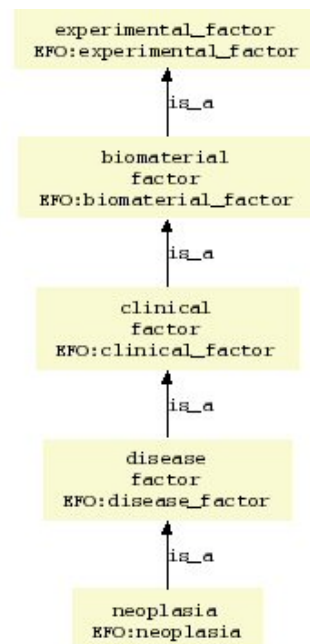


Figure 3 EFO term ‘neoplasia’ visualized in OLS

to assess granularity and overall coverage of the ontology and structure is manually evaluated by the curators who use the ontology.

Version 0.1 of EFO produces automated mappings comparable in coverage (~35%) for a 6000 sample test set between the EFO and the NCI thesaurus. Replacing the NCI thesaurus with the EFO reduced false positives and multiple matches by an order of magnitude (60% reduced to 8.6%). We believe by continuing an iterative process of mapping, curating and integrating EFO terms alongside an iterative evaluation strategy and restructuring the ontology we can continue to improve the quality and coverage of the ontology throughout its lifecycle.

## 4 DISCUSSION

It is our belief that application ontologies such as the EFO should be constructed with a principal to minimize redundancy and maximize information sharing. Wherever possible, mapping to external resources such as OBO Foundry ontologies increases interoperability through a common and shared understanding. Furthermore, this removes the temptation to ‘reinvent the wheel’ and allows the exploitation of the efforts currently underway to represent particular communities. It also permits updating when reference ontologies change.

A complication of this approach is the implication of mapping to external ontology concepts and their implicit hierarchy. In EFO our ‘meaning’ is limited to the textual definitions of the concepts externally mapped to EFO terms. Importing and accepting all axioms associated with concepts is a desirable long term goal. However the potential for conflicting logical definitions and lack of an intuitive standardized and easy to use upper ontology framework have caused us to initially defer this task. BFO (Grenon *et al.*, 2004) was not considered as an upper level ontology for EFO in its earliest form as the primary focus of this project is the application of the ontology and rapid development. However, mapping to BFO (or some other upper level ontology) is something we are now beginning to look into for future development and will appear in the forthcoming future releases.

The OBO Foundry has resolved issues, of orthogonal coverage and unique namespace identifiers and has made our task easier. In the future we will make bimonthly releases of EFO, continue the validation process, consider requests for new terms and map additional data resources to the EFO. GEO data imported into the ArrayExpress framework is already mapped during import, and any data resource with biological annotation could be mapped semi-automatically. Obvious candidates include Uniprot and other gene expression databases which are targets for integration with ArrayExpress. Version 0.2 of EFO is available from the EBI Ontology Lookup Service,

comments and questions can be sent to [exfactorontology@ebi.ac.uk](mailto:exfactorontology@ebi.ac.uk)

## ACKNOWLEDGEMENTS

The authors are funded in part by EC grants FELICS (contract number 021902), EMERALD (project number LSHG-CT-2006-037686), Gen2Phen (contract number 200754) and by EMBL. Thanks to the ArrayExpress production team: Anna Farne, Ele Holloway, Margus Lukk, and Eleanor Williams for useful comments.

## REFERENCES

- Bard, J, Rhee, SY *et al.* (2005) *An ontology for cell types. Genome Biol.* 6(2): R21.
- Bard, J, and Rhee, SY (2004) *Ontologies in biology: design, applications and future challenges. Nature Rev Gen* 5, 213-222.
- Brazma, A, Hingamp, P *et al.* (2001) *Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nature Genetics* 29, 365-371.
- Brinkley, JF, Suci, D *et al.* (2006) *A framework for using reference ontologies as a foundation for the semantic web. In Proceedings, American Medical Informatics Association Fall Symposium, 96-100, Bethesda, MD.*
- Dyck, P and Chisholm, R (2003) *Disease Ontology: Structuring Medical Billing Codes for Medical Record Mining and Disease Gene Association. Proceedings of the Sixth Annual Bio-ontologies Meeting, Brisbane, 2003, 53-55.*
- Fragoso, G, de Coronado, S, *et al* (2004) *Overview and Utilization of the NCI Thesaurus. Comp Func Gen* 5:8:648-654.
- Grenon, P., Smith, B. *et al.* (2004) *Biodynamic ontology: applying BFO in the biomedical domain. In Ontologies in Medicine (ed. Pisanelli, D.M.) 20–38 (IOS, Amsterdam, 2004).*
- Levenshtein, VI (1966) *Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady* 10:707–710.
- Parkinson H, Kapushesky M, *et al.* (2006). *ArrayExpress - a public database of microarray experiments and gene expression profiles. Nucl Acids Res* 35, D747-750.
- Phillips L (1990) *Hanging on the Metaphone. Comp Lan* 7: 39-49.
- Phillips L (2000) *The Double Metaphone Search Algorithm. C/C++ Users Journal.*
- Schober, D, Kusnierczyk, W *et al.* (2007) *Towards naming conventions for use in controlled vocabulary and ontology engineering, Proceedings of the Bio-Ontologies Workshop, ISMB/ECCB, Vienna, July 20, 2007, 87-90.*
- Smith B, Ashburner, M, *et al.* (2007). *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol* 25, 1251.
- Sprague, J, Bayraktaroglu L, (2006) *The Zebrafish Information Network: the zebrafish model organism database. Nucleic Acids Res.* 34, D581–D585.
- Uschold, M, Grüninger, M (1996) *Ontology: Principles, Methods and Applications. Knowledge Engineering Review* 11(2), 93-155.