

Logical axiomatization of the Evidence & Conclusion Ontology (ECO) by integrating external ontology classes

Rebecca Tauber¹ & Marcus C. Chibucus^{1,2*}

¹Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland, 21201 USA

²Department of Microbiology and Immunology

*correspondence: mchibucus@som.umaryland.edu

ABSTRACT

Mapping semantically equivalent classes across ontologies is a crucial step toward increasing interoperability and is necessary to enable the leveraging of existing external ontologies during ontology development. Interoperability can allow the adoption of logical design patterns, which can enhance ontology manageability, improve structural consistency, and reduce development time, in addition to facilitating knowledge discovery. The Evidence & Conclusion Ontology (ECO) and the Ontology for Biomedical Investigations (OBI) began a loose collaboration, i.e. talking, in 2011. Recently, however, great strides have been made toward harmonizing these two ontologies through integrating components of OBI into ECO, i.e. creating logical definitions in ECO using imported OBI classes. As these are two orthogonal OWL ontologies, enabling such integration required creation of a logical design pattern to transform OBI classes (which define instruments, assays, etc.) into equivalent ECO evidence classes. This design pattern allows ECO to harness the expressivity of OBI in capturing complex experimental workflows that generate “evidence” that is cited in scientific publications. The goals of this effort are to increase consistency in the structure of ECO, facilitate further ECO and OBI development, better describe the methodologies that produce evidence, and discover new relationships between ECO evidence types. Here, we present the methods for integration and discuss this work as a model for future ontology harmonization efforts.

1 STRUCTURING SCIENTIFIC EVIDENCE

When interpreting the findings of any scientific investigation, “evidence” is an important aspect to consider. What methods were employed? What types of data were generated? How were findings interpreted? Documenting aspects of the scientific methodology employed in a given study affords investigators a basis for interpreting the results.

Ultimately researchers use evidence to support a variety of conclusions. In the biomedical realm, one such conclusion might be the interpretation that a protein has a particular function. Professional biocurators meticulously extract such information – about methods, evidence, and conclusions – from the scientific literature using a variety of manual and automated methods. This information is represented variously so that it can be stored at databases where it can be readily manipulated and used by researchers.

1.1 The Evidence & Conclusion Ontology

The Evidence & Conclusion Ontology (ECO)¹ systematically describes types of scientific evidence in biological research, such as evidence generated from laboratory experiments, computational methods, or statements curated from

Funding acknowledgement: This material is based upon work supported by the National Science Foundation (NSF) Division of Biological Infrastructure (DBI) under Award Number 1458400 to MCC.

literature (**Fig. 1**). ECO represents a range of evidence categories spanning from broad (e.g. ‘sequence similarity evidence’ or ‘author statement evidence’) to specific (e.g. ‘sodium dodecyl sulfate polyacrylamide gel electrophoresis evidence’). Evidence types, as summarized by over 800 ECO classes, become important pieces of metadata associated with annotations at databases that are used by researchers worldwide to support their investigations.

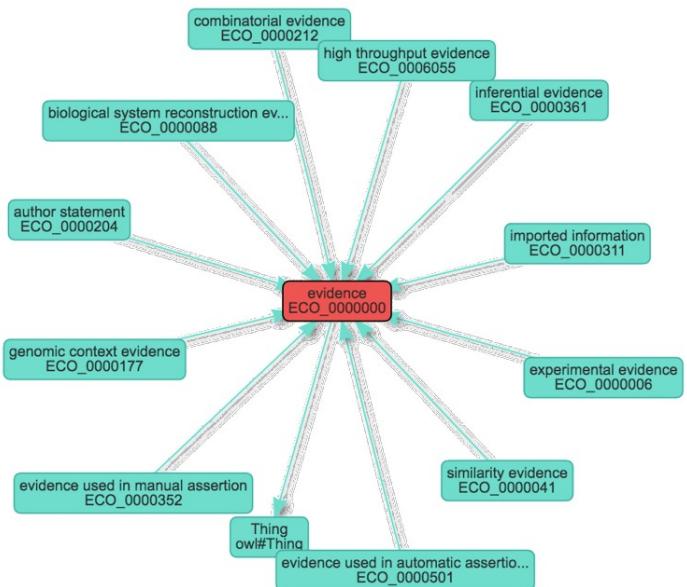


Fig. 1. ECO’s current highest-level evidence classes as depicted at <http://evidenceontology.org/browse>

ECO terms, as ontology classes, contain standard definitions and synonyms and are networked with relationships. Thus, associating research data with ECO evidence terms allows bioinformatics resources to manage large volumes of annotation data by providing mechanisms for sorting, querying, and performing quality control checks. For example, UniProt-Gene Ontology Annotation (UniProt-GOA) uses ECO to support searching of more than 365 million evidence-linked GO annotations² and the Gene Ontology³ resource itself uses ECO in support of various quality control mechanisms including annotation consistency.⁴

1.1.1. Axes of classification

The axes of classification in ECO are ‘evidence’ and ‘assertion method’, which are disjoint from one another (**Fig. 2**).

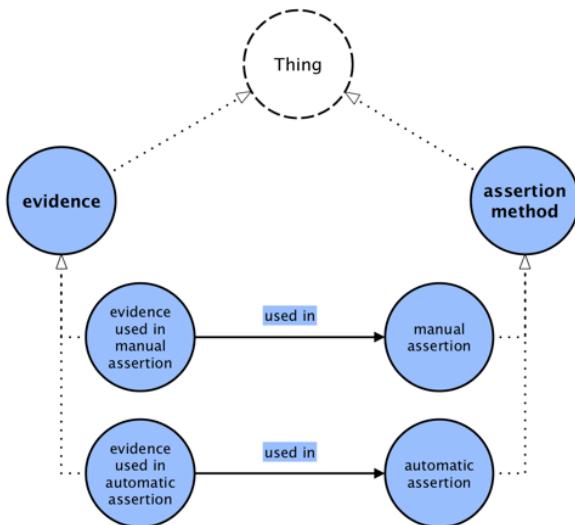


Fig. 2. The two ECO root classes.

‘evidence’ (ECO:0000000)

‘Evidence’ (ECO:0000000) - defined as “a type of information that is used to support an assertion” - can be thought of as a description that may be representative of both the broad methods employed and any outputs generated by such methods. For example, ‘clinical study evidence’ (ECO:0000180) may refer both to the protocols used and types of data generated during a controlled investigation that uses human subjects.

Consider ‘chromatography evidence’ (ECO:0000325), which is defined as “a type of experimental evidence that is based on separation of constituent parts of a mixture (the mobile phase) as they pass differentially through a stationary phase due to differences in partition coefficient and retention on the stationary phase.” A researcher considering some scientific conclusion supported by chromatography evidence might be evaluating a graph generated during a chromatography experiment that depicts a peak, which represents light absorbance and elution time from a stationary column. But the peak alone is not taken as the evidence: the results are considered within a particular context. Experimental conditions such as the type of solvent or column used or observations such as how the chromatograph peak compares to peaks made with known standards are considered, as well.

Thus, ECO classes are considered summary in nature. Each class can be seen as a type of ‘information content entity’ (IAO:0000030) from Information Artifact Ontology

(IAO)⁵, which defines ‘information content entity’ as “a generically dependent continuant that is about some thing.”

‘assertion method’ (ECO:0000217)

‘Assertion method’ (ECO:0000217) is the second root class of ECO in addition to ‘evidence’, and it is used to describe whether a human being (e.g. a professional biocurator) or a machine (e.g. a computational pipeline) generated a particular evidence-based annotation that is stored at a biological database. This class and its node within the ECO ontology have a complex history outside the present discussion (see Chibucos, et al. 2014¹ for a more thorough discussion). Briefly, ‘assertion method’ has only two subclasses, ‘manual assertion’ and ‘automatic assertion’, which refer to statements made by humans and machines, respectively.

Connecting ‘evidence’ and ‘assertion method’

‘Evidence’ is logically tied to ‘assertion method’ through the ‘used in’ relationship, enabling one to state whether a person or machine applied a particular piece of evidence in making an annotation (**Fig. 2**). For example, a human biocurator reading the literature to generate biological database annotations might read a scientific article where some ‘experimental evidence’ (ECO:0000006) was presented about some metabolic pathway and its association with some disease in some organism. After carefully interpreting the methods and results presented in the paper, the biocurator might draw a conclusion such as “metabolic pathway x is involved in disease y”.

This conclusion might be *asserted* by the curator, typically as a database annotation that could include multiple other pieces of information, depending on the database. Because a person made the annotation, i.e. ‘manual assertion’ (ECO:0000218), and the evidence supporting the annotation was ‘experimental evidence’, these two disjoint classes become connected as ‘experimental evidence used in manual assertion’ (ECO:0000269).

Simultaneously recording both ‘evidence’ and ‘assertion method’ gives databases another dimension for interpreting and presenting data. (Note: the ‘used in’ relationship is under review and this structure of ECO is subject to continued development.)

1.1.2. Current ECO status

As ECO’s user base has continued to grow, so has the number of classes. As of July 2017, there were 513 *pure* ‘evidence’ classes, i.e. those not linked logically to ‘assertion method’ but which have a subclass that is so linked. 316 additional classes were of the ‘used in manual assertion’ type, meaning that they are children of one of the approximately 500 pure evidence classes, combined with the ‘used in’ logical definition for a ‘manual assertion’. Finally, there were 54 ‘used in automatic assertion’ terms.

Up to this point, ECO has primarily been a class hierarchy, only utilizing a ‘used in’ property to logically define how the evidence was generated. The addition of more logical definitions through incorporation of the Ontology for Biomedical Investigations (OBI)⁶ can lead to discovery of new relationships through reasoning and facilitate development speed & consistency. It has also helped to further clarify ECO’s axes of classification and standardize ECO’s English definitions.

2 INTRODUCTION TO OBI

The Ontology for Biomedical Investigations (OBI)⁶ describes scientific investigations, e.g. study design & execution, instruments & processes, data analysis, and so on, and can be used to model how aspects of an investigation interrelate. OBI, like ECO, is developed in Web Ontology Language (OWL). OBI uses upper-level Basic Formal Ontology (BFO)⁷ classes to guide development. BFO top-level classes include ‘continuant’ and ‘occurrent’.

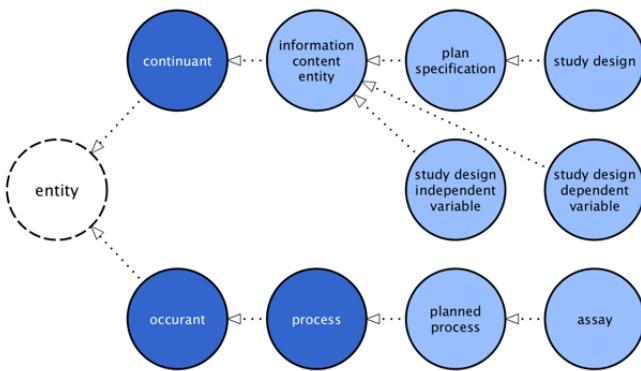


Fig. 3. Selected BFO⁷ classes (dark blue) and OBI⁶ classes (light blue).

OBI uses logical axioms to describe different parts of biomedical investigations, which allows for very detailed modeling of such investigations. As shown in **Fig. 3**, the parts of an investigation may include a study design, independent and dependent variables, and the assay conducted. These are important components of the ECO-OBI mappings.

3 MAPPING ONTOLOGY CLASSES

In order to make use of the logic already inherent to OBI, ECO classes must be mapped to their equivalent OBI class(es), which already utilize various logical definitions. The mappings import that logic to be used in reasoning for structural analysis and future knowledge discovery. Not only can the ECO structure be reviewed and revised, but also these mappings provide benefits to ECO users who

annotate evidence from complex workflows (and would like to see a tidy summary class).

Ideally, mapping classes between ontologies can be a straightforward process. A class axiom using the owl:equivalentClass property is added to link a class in one ontology to an equivalent class in another. However, this is only possible and logically correct between heterogeneous ontologies. In the case of orthogonal ontologies, it is easy to see a correlation between two terms, but it is much more difficult to transform this into a class axiom. For example, while ECO may define ‘microscopy evidence’, OBI defines the process of ‘microscopy’. How does one state that the process of microscopy results in microscopy evidence?

To make this logical transformation, an alignment Ontology Design Pattern (ODP) must be created. This serves as an OWL template to be inserted as the object of the equivalence class axiom. In reality, even the simple axiom ‘x owl:equivalentClass y’ is an ODP, but, out of necessity, the ODPS for orthogonal ontologies tend to be more complex.

3.1 Ontology Design Pattern

The ECO-OBI ODP consists of four distinct components that are combined to create the mappings (**Fig. 4**). The % symbol is replaced with the OBI class for the mapping, and ‘evidence’ is replaced with the direct parent of the evidence class being mapped. These axioms are either equivalence or subclass statements, depending on the degree of specificity that can be achieved with existing OBI classes.

- The study design, which specifies the plan to generate evidence
`evidence and is_specified_output_of some ('study design execution' and (realizes some (concretizes some %)))`
- The assay, which is the process that generates evidence
`evidence and is_specified_output_of some %`
- The independent variable, which is manipulated to generate evidence
`evidence and is_specified_output_of some ('study design execution' and (realizes some (concretizes some ('study design' and 'has part' some 'study design independent variable' and 'is about' some %))))`
- The dependent variable, which the evidence is about
`evidence and is_specified_output_of some ('study design execution' and (realizes some (concretizes some ('study design' and 'has part' some 'study design dependent variable' and 'is about' some %))))`

Fig. 4. ECO-OBI Ontology Design Pattern (ODP) components with OWL axiom (blue text)

It is important to note that each mapping may use anywhere from one to all of the components, depending on the complexity of the processes involved in generating the evidence. Specifically, many ECO evidence classes may not

include an independent variable that has been manipulated to assess a dependent variable. This is true for assays that measure, detect, prepare, or simply visualize specimens, such as microscopy.

Many classes have completed mappings to all four ODP components (**Fig. 5**).

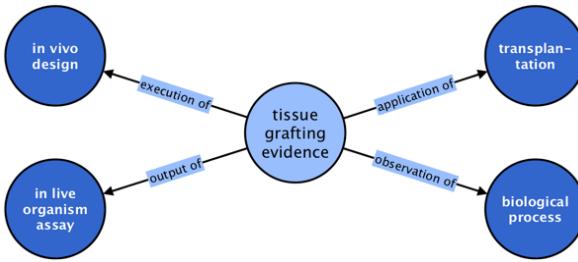


Fig. 5. ECO class (light blue) with completed mapping to OBI (dark blue). This particular mapping is a subclass statement, as there are no OBI classes specific enough to make an equivalence axiom logically correct.

The actual subclass statement for ‘tissue grafting evidence’ utilizes all four ECO mappings to OBI (**Fig. 6**).

```

anatomical perturbation evidence
and is_specified_output_of some ('study design execution'
    and (realizes some (concretizes some 'in vivo design'))) # study design
and is_specified_output_of some 'in live organism assay' # assay
and is_specified_output_of some ('study design execution'
    and (realizes some (concretizes some ('study design'
        and 'has part' some 'study design independent variable'
        and 'is about' some transplantation)))) # independent var.
and is_specified_output_of some ('study design execution'
    and (realizes some (concretizes some ('study design'
        and 'has part' some 'study design dependent variable'
        and 'is about' some biological_process)))) # dependent var.
  
```

Fig. 6. Subclass statement relating ECO ‘tissue grafting evidence’ to OBI

3.2 Mapping process

Before any mappings could begin, we needed to retrieve a set of ECO classes for testing. The ‘experimental evidence’ node of ECO was chosen because these evidence classes can more easily be associated with various assays found in OBI. A SPARQL query was performed to get all children of ‘experimental evidence’ and associated axioms as a CSV.

In order to facilitate the workflow, the CSV was exported to a Google spreadsheet and headers were added with space for each component of the design pattern. This way,

we were able to go through, row by row, and determine the best fit for each. This required manual review of the ECO class and manual searches of both OBI and GO. After we determined the design pattern was feasible, it was time to test the axioms in the ontology itself.

ROBOT⁸ is a versatile tool for working with OWL ontologies and was created to work with biomedical ontologies, although it can easily be applied to any ontology development. It allows developers to perform a variety of tasks, from filtering, to merging, and even converting ontology formats. One of the most useful features of ROBOT (and the one that was utilized for our harmonization efforts) is the template command. The spreadsheet created in the previous step was formatted with specific headers that ROBOT uses to transform the cell contents into axioms. The ROBOT template we used is demonstrated in **Table 1**, with two examples of mappings.

ECO ID	Class Type	ECO Parent	OBI Study Design	OBI Independent Variable	OBI Dependent Variable	OBI Assay
ID	CLASS_TYPE	C	C ...	C ...	C ...	C ...
ECO:0000243 (tissue grafting evidence)	subclass	anatomical perturbation evidence	in vivo design	transplantation	biological process	in live organism assay
ECO:0001062 (in vitro demethylation assay evidence)	equivalent	in vitro assay evidence	in vitro design		demethylation	in container assay

Table 1. ROBOT template displaying mapping between two ECO classes and respective OBI classes (explained in text).

As shown in **Table 1**, the first row contains human readable labels for each column that are not parsed by ROBOT. The second row contains the template strings. If a cell in the second row begins with a ‘C’, all entries in that column will be parsed as logical axioms. On the other hand, if it were to begin with an ‘A’, it would be parsed as an annotation. For the OBI columns, the ‘...’ in row two contains the OWL axioms shown in the design pattern, and the % symbol is replaced by the content in a given cell. The column ‘CLASS_TYPE’ specifies if the generated axiom is either a type of subclass or equivalent statement.

After populating the table, for each ECO class in the ID column of a ROBOT template, ROBOT will parse the contents of that row and build an axiom based on the information in each cell that corresponds to the template strings in the column headers.

3.2.1. Results of mapping ECO-OBI

The axioms created by the ROBOT template were immediately merged into ECO and reviewed in Protégé.

Throughout the mapping process, we detected areas of OBI to expand. In some cases, OBI did not have enough terms to create an accurate mapping, so term suggestions were made. We are currently in the process of requesting the addition of 40 assay classes and 24 non-assay classes. Once these new terms have been accepted into OBI, 161 mappings using them will be added to the ECO working branch on GitHub⁹ for review.

We believe that expending the effort to map ECO and OBI has already been worth the effort. It has identified areas for OBI development, resulted in greater logic within ECO, and helped disentangle confused axes of classification within ECO. Work will continue on harmonizing ECO and OBI using the experimental node of ECO initially but expanding eventually to other areas, e.g. sequence similarity.

After ECO and OBI have robust mappings, we believe that eventually ECO can leverage other external ontologies in a similar fashion.

ACKNOWLEDGEMENTS

Special thanks to: Elvira Mitraka for term review assistance; Matthew Brush for contributing the ODP, which was conceived during a 2016 joint OBI-ECO meeting on evidence in Baltimore; and Bjoern Peters, James Overton, Christian J. Stoeckert, Jr. and other OBI developers for collaborating.

REFERENCES

1. Chibucos, M.C., Mungall, C.J., Balakrishnan, R., Christie, K.R., Huntley, R.P., White, O., Blake, J.A., Lewis, S.E., and Giglio, M. (2014) Standardized description of scientific evidence using the Evidence Ontology (ECO). *Database (Oxford)*, v.2014:bau075.
2. Dimmer, E.C., Huntley, R.P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martinet, M.J., et al. (2012) The UniProt-GO Annotation database in 2011. *Nucleic Acids Research*. **40**:D565–D570.
3. The Gene Ontology Consortium. (2016) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*. **45**(D1):D331-D338.
4. Chibucos, M.C., Siegela, D.A., Hu, J.C., Giglio, M. (2017) The Evidence and Conclusion Ontology (ECO): Supporting GO Annotations. In Christophe Dessimoz & Nives Škunca (eds.), *The Gene Ontology Handbook*, Methods in Molecular Biology, vol. 1446, pp. 245-259. New York City: Humana Press (Springer). ISBN 978-1-4939-3743-1
5. <https://github.com/information-artifact-ontology/IAO>
6. Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M.H., Bug, B., Chibucos, M.C., et al., (2016) The Ontology for Biomedical Investigations. *PLoS One*. **11**(4):e0154556.
7. Arp, R., Smith, B., Spear, A.D. (2015) Building Ontologies with Basic Formal Ontology. Cambridge: The MIT Press.
8. ROBOT on GitHub: <https://github.com/ontodev/robot>
9. ECO on GitHub: <https://github.com/evidenceontology/evidenceontology>