GO-CAM Specification Proposal (v 0.7)

Note the users should refer to the version of this document on the GO wiki: http://wiki.geneontology.org/index.php/Noctua

Note: Work in progress. This document is meant to be a more descriptive version of the formal specifications, written in ShEx format (https://github.com/geneontology/GO_Shapes/blob/master/shapes/go-cam-shapes.shex) The specifications are not final.

Note: This document describes the abstract data model for the GO-CAMs. It provides instructions to develop tools for GO-CAM models. Note that this document is still a work in progress and not the final specification.

Graphical summary of the GO-CAM Specification Proposal v0.7

Section 1. Activity Unit
- Components of an activity unit
  - Molecular activity (GO MF)
  - Active entity
  - Biological process (BP)
  - Location
  - Spatial context
  - Temporal context
- Relationships relevant to the Activity Unit
  - enabled_by/(contributes to)
  - has_input RO:xxx
  - has_output
  - occurs_in
  - occurent_part_of
  - happens_during

Section 2. Causal effect relations
- Effect of one activity unit on another molecular activity
  - Directly positively regulates
  - Directly negatively regulates
  - Directly regulates
  - Causally upstream of, positive effect

Causally upstream of, negative effect
Causally upstream of
Effect of one activity on another activity, that is dependent on a constitutive “activity-regulating” process
Effect of an activity on a larger biological process
Positively regulates
Negatively regulates
Regulates
Causally upstream of
Causally upstream of, positive effect
Causally upstream of, negative effect

Section 3. Isolated (unconnected) GO-CAM statements
Observation of subcellular localization of a gene product
Observation of an effect of a gene product perturbation on a larger process (usually from phenotype)
NOT annotations

discuss
Shared BP
Causally upstream of or within

Future developments
Combining effect relations (NOTE: this is forward thinking, so it should not be part of the official spec)

Contributors

Glossary

Graphical summary of the GO-CAM Specification Proposal v0.7

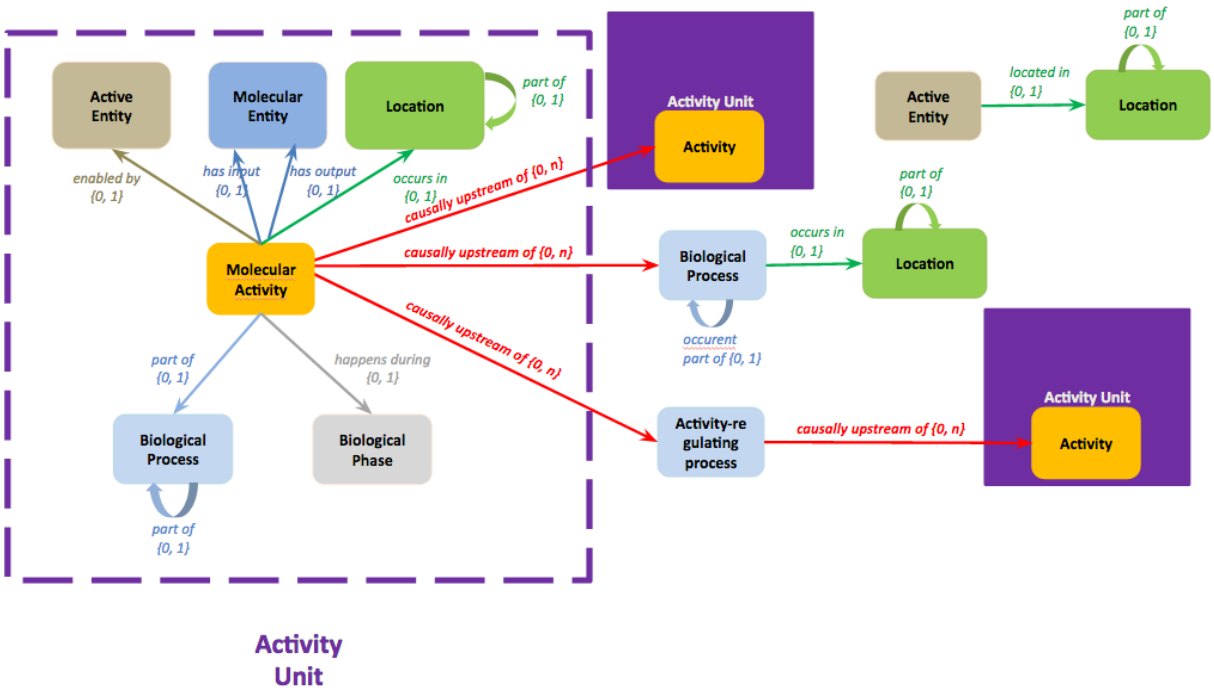


Figure 1. Summary of GO-CAM model specifications. The left-hand panel represents a detailed *activity unit* (see Section 1) and the 3 ways to connect them together (see Section 2). The upper-right *active entity* located_in location represents independent statements about gene product locations.

Section 1. Activity Unit

Definition: An *activity unit* is centered around a molecular activity (GO Molecular Function), and connects this activity to the cellular component in which it occurs, the biological process it is part of, and the spatial and temporal context. An *activity unit* minimally contains one *Molecular Activity.*

The *activity unit* can be related to another activity unit via a *causal relation* between the molecular activities of each one (MF to MF effects). It can also be related via a *causal relation* to larger biological processes (MF to BP effects).

Components of an activity unit

An *activity unit* is composed of the following elements (Figure 2): *active entity, molecular function, target entity, cellular component, biological process, location* and *temporal context.* Below is a description of each of these classes.

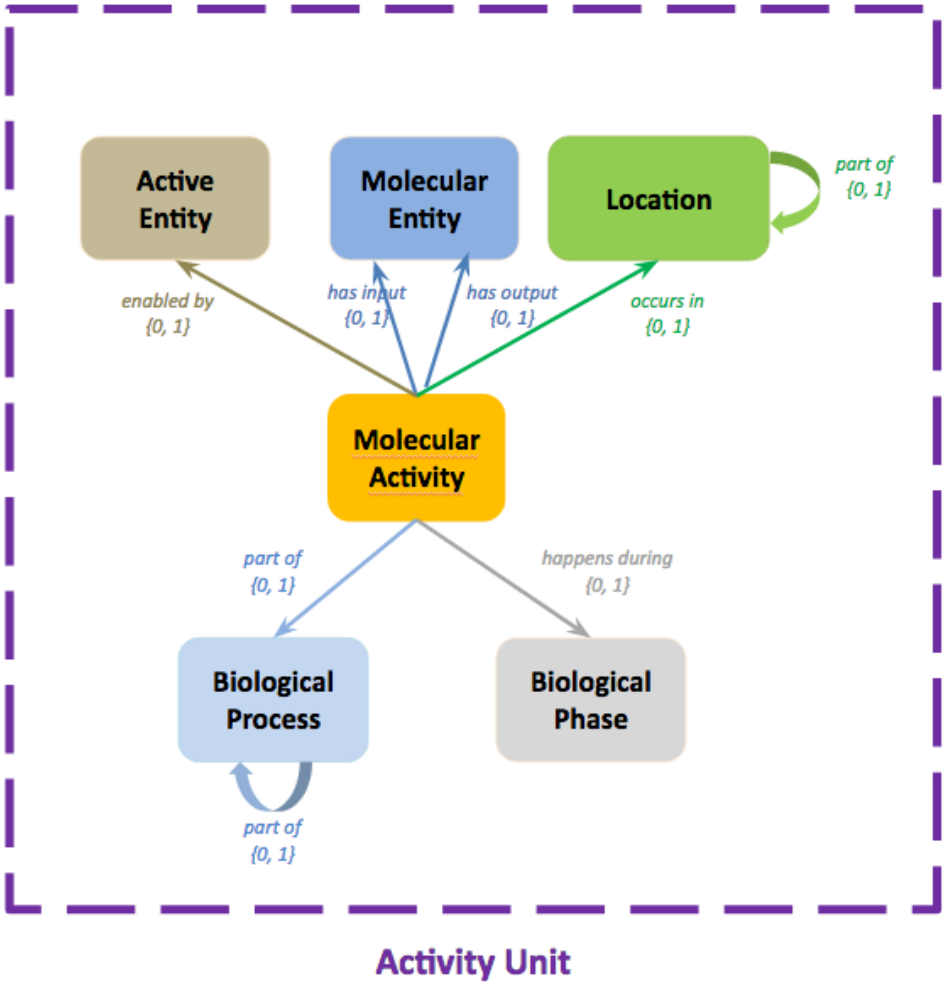


Figure 2. An *activity unit* comprises all the elements contained in the dotted box. The *activity unit* represents a molecular function that is part of some biological process and is enabled by an *active entity* (gene product or complex); it occurs in some location, happens during some biological phase and could require some target entity as input, and generate an output.

Table 1. GO-CAM elements and ontologies used. Note that the formalism follows GO annotation practice: gene products (or complexes comprised of multiple gene products) have molecular activities (GO molecular function), are active in specific locations (GO cellular component) and act as part of larger biological programs (GO biological process). Other elements of GO-CAM provide further structured extensions of standard GO annotations.

GO-CAM element (Figure 2)	**Ontology or identifier source(s)**	**Example**
Molecular activity	GO molecular function	ubiquitin-protein transferase activity (GO:0004842)
Biological process	GO biological process	cellular response to UV (GO:0034644)

Location	GO cellular component	nucleus (GO:0005634)
	Cell Type Ontology (CL) (8)	retinal cell (CL: 0009004)
	anatomy ontologies, e.g. UBERON (9), *C. elegans* gross anatomy (10), EMAPA (11)	eye (UBERON: 0000970)
Active entity	Gene, protein, RNA or complex identifier from a standard source, e.g. HGNC for a human gene	NEDD4 (HGNC:7727)
Target entity	Same as active entity, or chemical from ChEBI (12)	MAP2K1 (HGNC:6840)
Biological phase	GO biological phase (GO:0044848)	mitotic G1 phase (GO:0000080)
	Developmental phase ontology, e.g. Mouse Developmental Stage	Theiler stage 02 (MmusDv:0000005)
Relations (arrows in Figure 2)	Relations Ontology	*occurs in* (BFO:0000066)

Molecular activity (GO MF)

Definition: **This is the central element of the *activity unit*.** This is the GO MF carried out by the *active entity.* It is an occurrent.
Usage:

- Each *activity unit* must contain **exactly** one MF (use root MF term when the activity is unknown). When a gene product has more than one distinct MF, each function should be in a separate *activity unit*. The *activity* is related to the *active entity* via the **enabled_by** relation.
- When the root MF term is assigned, it means the function is unknown or unspecified. This can be used to indicate that an a*ctive entity* is part of (or causally upstream of) a BP while its *activity* is not specified.

Active entity

Definition: A molecular machine present in an active conformation with enough quantity to carry out the MF in vivo. It is a continuant. The active entity can be a gene or gene product, a gene product-containing complex. It can also be generalized to a SET, where each member of the set is capable of the given activity. When defining a set, the following logic operators should be used to indicate their relationship:

- AND: all entities are required for the activity, ie an emergent function of the complex;
- OR: either entity is required for the activity

Sets are instantiated as active entities in OWL. The instance-based model used by GO-CAMs does not have the ability to represent logical functions such as unions or intersections. When these are required, we refer to a logically defined class in an ontology.

Usage:

- The *active entity* can be 0 (when it is unknown or unspecified) or one (protein complexes are modeled as a single entity unit and considered one node).

Molecular entity

Definition: This is either the molecule that is acted upon by the , or the molecule that is produced by the MF of the *active entity*..molecular function.
Usage:

- In general, the *input entity* does not need to be specified, as it may be implied by either the MF, or the active entity of the downstream *activity unit*. For example, if the MF is acetylcholinesterase activity, the input is acetylcholine. Inputs should be specified by the ontology.
- The main use case is to specify a protein target, when that protein's activity is not regulated. For example, if a protein is phosphorylated, making it accessible to become the substrate in a subsequent reaction, it should be specified with **has_input**.
- A special case occurs when the MF is the GO term "protein binding". In this case, it is possible to use multiple **has_input** relations to specify the binding partners, and no **enabled_by**, as the binding reaction can be considered symmetrical (this aligns with the GO annotation rule of reciprocal protein binding annotations).The products of reactions are specified with the **has_output** relation.

Biological process (BP)

Definition: The BP (biological program) that the molecular function is part of. It is an occurrent.
Usage:

- Each *activity unit* can contain 0 or one GO BP.
- BP is related to the MF via the **part_of** relation.
- Other processes can be included. These must be nested via **part_of** relations.

- It is possible to associate an entity with a BP without knowing what the exact molecular activity is.

Location

Definition: The location of the *active entity* when it executes the molecular function.
Usage:

- Each *activity unit* can have 0 or one location.
- It is related to the MF via the **occurs_in** relation.
- The location can be specified at multiple levels of biological organization. When multiple levels are specified, the MF must be connected to the lowest level via the occurs in relation, which is nested within one or more higher levels by using the part of relation, e.g. MF occurs in nucleus (GO CC) part of heart (anatomy). The levels are as follows, from lowest to highest. Only one part of relation can be made from one level to the next:
    - GO cellular component (note that there can be multiple nested levels of GO cellular component)
    - Cell type
    - Anatomy
    - Organism
-

Spatial context

Definition: A GO CC of the *activity unit* can be nested within other ontology terms describing further spatial context. These ontologies include Cell Type, Tissue type and/or Anatomical Ontologies. For example, the cellular component may belong to a particular cell type.
Usage:

- The spatial context is related to the MF via the **part_of** relation. If the CC is unknown or unspecified, the spatial context can be linked directly to the MF.
- Spatial context can be further modified in cases where a structure is adjacent to one of 2 cellular components, or overlaps cells or cellular components+++
  adjacent_to and **overlaps**

Temporal context

Definition: A GO MF can occur at a specified temporal phase, either a GO biological phase (e.g. G1 phase of the cell cycle), or a developmental stage, e.g. Mouse Theiler stage.
Usage:

- The temporal context is related to the MF via the using the **happens_during** relation.

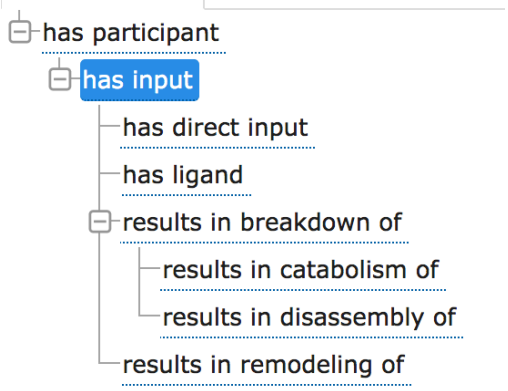
Relationships relevant to the Activity Unit

enabled_by/(contributes to)

Confirm usage of contributes to. As a GPAD export relation only? What about the MOD imports?
Domain (subject): GO:0003674 Molecular Function and children
Local range (object): CHEBI:33695 information biomacromolecule (gene or gene product)

has_input RO:xxx

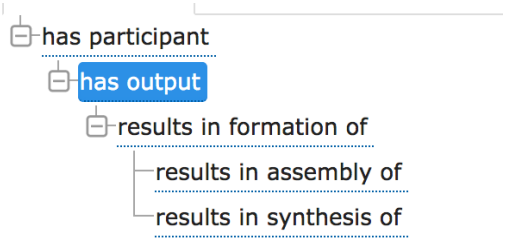
No other child term of has_input will be allowed in GO-CAM models.
Make sure we include domain and range for this.
Domain (subject): GO:0003674 Molecular Function and children
Local range (object): CHEBI:24431 chemical entity (including CHEBI:33695 information biomacromolecule (gene or gene product))



has_output

Not in the diagram; note that this term has children
Should add to the spec, but need curation guidelines

Domain (subject): GO:0003824 catalytic activity
Local range (object): CHEBI:24431 chemical entity (including CHEBI:33695 information biomacromolecule (gene or gene product))



occurs_in

Domain (subject): GO:0003674 Molecular Function and children
Local range (object): CL:0000000 GO:0005575 PO:0025131 UBERON:0001062 WBbt:0004017 WBbt:0005766 NCBITaxon:1

adjacent_to

Domain (subject): cellular components that are not part_of a cell (for eg. extracellular region, extracellular space and extracellular matrix, possibly add more on a need basis)
Local range (object): CL:0000000, WBbt:0004017 Cell (etc?)
Cardinality: 0,1

overlaps

Domain (subject): cellular components that includes parts from more than one cell (for eg. GO:0031594 'neuromuscular junction synapse). As a first list of allowed terms, maybe we can use those using 'overlaps' in the logical definitions and subclass
Local range (object): CL:0000000, WBbt:0004017 Cell (etc?)
Cardinality: 0,2,3 (3 = GO:0061689 'tricellular tight junction')
GO:0031594 'neuromuscular junction synapse : (overlaps some 'motor neuron') and (overlaps some 'muscle cell')

happens_during

Domain (subject): GO:0003674 Molecular Function and children
Local range: GO:0008150 WBls:0000075 ZFS:0100000 UBERON:0000105

Section 2. Causal effect relations

Causal relations describe how an *activity unit* affects either another *activity unit*, or the outcome of a larger biological process.

Effect of one activity unit on another molecular activity

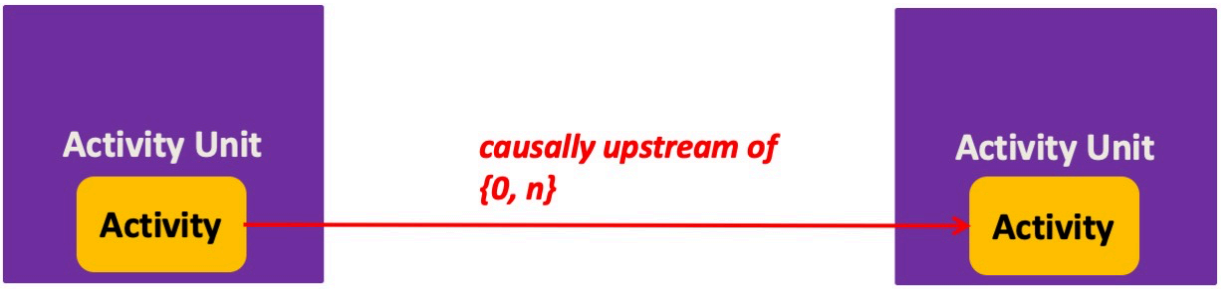


Figure 3. How *activity units* connect together. The first and simplest case is to link two activity units together through their activities. The second case shows an activity unit causally upstream of a biological process that occurs in some location. The third way to connect activities is through an "activity-regulating process" (certain types of Biological Processes) that regulate the activity levels of some other activities.

Definition: The effect exerted by one *activity unit* on another *activity unit*. It mainly represents the same semantics as the "regulation of molecular function" classes currently in GO. The types of effects are specified by the following relations between the molecular function of each *activity unit*.

Directly positively regulates

Definition: A direct action, usually involving direct protein-protein interaction, that produces a positive or activating effect from one *activity unit* to another.

Directly negatively regulates

Definition: A direct action, usually involving direct protein-protein interaction, that produces a negative or inhibiting effect from one *activity unit* to another.

Directly regulates

Definition: A direct action, usually involving direct protein-protein interaction, that produces either a neutral, or an unknown effect, from one *activity unit* to the other.

Causally upstream of, positive effect

Definition: An indirect (or unknown, i.e. not known to be direct) action that produces a positive or activating effect from one *activity unit* to the other.

Causally upstream of, negative effect

Definition: An indirect (or unknown, i.e. or not known to be direct) action that produces a negative or inhibiting effect from one *activity unit* to the other.

Causally upstream of

Definition: An indirect (or unknown, i.e. not known to be direct) action that produces either a *neutral* effect, or an effect of *unknown directionality*, from one *activity unit* to another.

Effect of one activity on another activity, that is dependent on a constitutive "activity-regulating" process

In some cases, the effect of one *activity unit* on another is mediated through a constitutive process that affects the amount of active gene product available in the correct conformation and location. We refer to these types of processes as '*activity-regulating processes*'. An example is the effect of a DNA-binding transcription factor on the activity of the product of the gene whose expression it regulates. Here, the activity-regulating process is transcription. In this case, we will have a model with the following pattern:

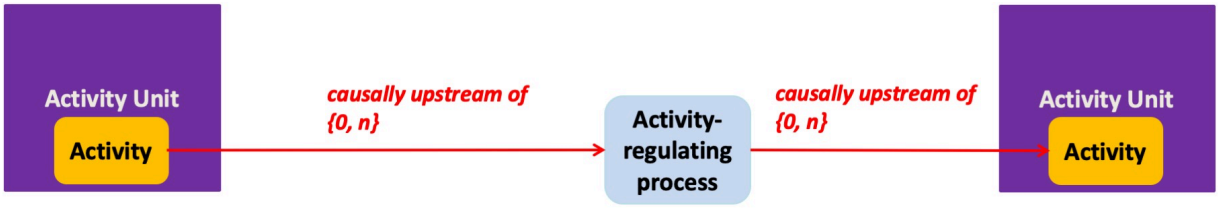


Figure 4. How *activity units* connect together via activity-regulating processes.

These "activity-regulating" BP's are restricted to just a few specific GO terms. This is the current list (it will likely be slightly expanded in the future):

- transcription, DNA templated (GO:0006351)
- ubiquitin-dependent protein catabolic process (GO:0006511)
- receptor internalization (GO:0031623)
- nuclear import (GO:0051170)
- translation (GO:0006412)
- (protein) sequestering
- (protein) secretion

Effect of an activity on a larger biological process

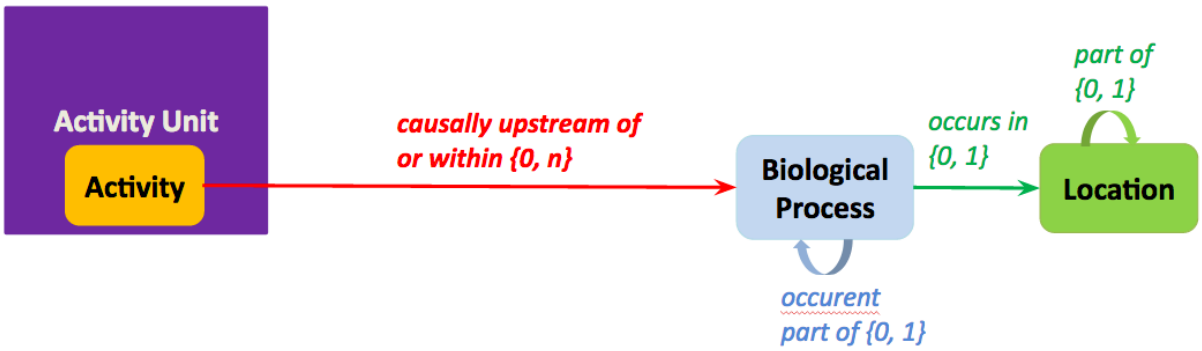


Figure 5. Effect of an activity on a larger biological process.

Definition: The effect exerted by an *activity unit* on a biological process. This relation should only be used if the activity unit is not a part of that biological process. It is used when an experiment measures a larger process (or some marker for that process). It mainly represents the same semantics as “regulation of biological process” classes in the GO, which are not themselves part of GO-CAM.

Positively regulates

Definition: An indirect action that produces a positive or activating effect of an *activity unit* on a biological process.

Negatively regulates

Definition: An indirect action that produces a negative or inhibiting effect of an *activity unit* on a biological process.

Regulates

Definition: An action, either direct or indirect, that produces a regulatory effect of one *activity unit* on a biological process. This relation does not specify whether the effect is positive or negative. It is used when the effect is either neutral, or of unknown direction.

Causally upstream of

Definition: This relation indicates that an *activity unit* precedes a biological process. It has no description about any regulatory action from one to the other. It can be used when the regulatory effect is unknown or irrelevant in the model.

Causally upstream of, positive effect

Definition: This relation indicates that an *activity unit* precedes a biological process. It has no description about any regulatory action from one to the other. It can be used when the effect direction is positive.

Causally upstream of, negative effect

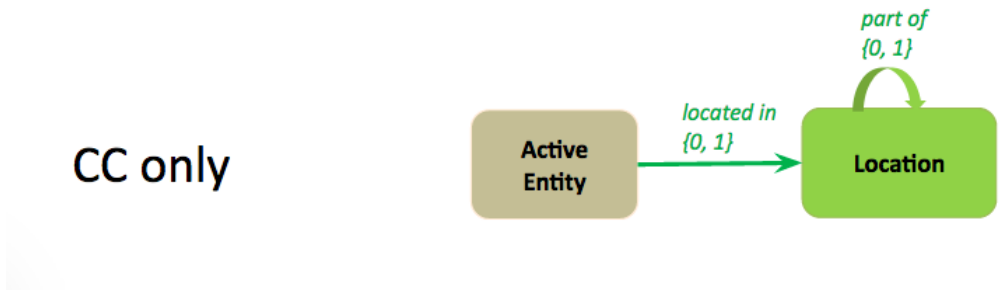
Definition: This relation indicates that an *activity unit* precedes a biological process. It has no description about any regulatory action from one to the other. It can be used when the effect direction is negative.

Section 3. Isolated (unconnected) GO-CAM statements

Some GO-CAM statements cannot be connected in the context of a larger model. However, they capture useful evidence that can be used in support of (or against) one or more triples in a larger model, and can be used to create standard GO annotations.

Observation of subcellular localization of a gene product

If a paper reports the subcellular localization of a gene product (GO Cellular Component), but it is not known whether the gene product is active in that location, the observation cannot necessarily be used in an activity unit. In this case, the information is represented as follows:



Observation of an effect of a gene product perturbation on a larger process (usually from phenotype)

If a paper reports the effect of a perturbation (e.g. knockout, over- or under-expression) of a gene product on a larger process, the observation cannot necessarily be used in an activity unit (as we do not know whether the gene product's activity is part of the process, or causally upstream of it). In this case, the information is represented as an isolated statement as follows: Active entity <--enabled by--- root MF ---causally_upstream_of_or_within→ BP

NOT annotations

If a paper supports a negative finding, this observation is not directly part of a GO-CAM model, but it is still useful for constraining one or more GO-CAM models.

discuss

- **Shared BP**
- **Causally upstream of or within**
    - Having part_of as a child of causally_upstream_of_or_within is ontology unsatisfactory - seems illogical.
    - One proposal is to use a boolean as a relation: ‘causally upstream’ OR ‘part of’ and have these as unconnected statements (roaming). To be connected, the relationship needs to be changed to either ‘causally upstream’ or ‘part of’.
-
- Paul proposes that we not allow BP occurs_in location triples, as location is captured at the level of molecular activities (GO annotates gene products) and the BP spans these locations

GO-CAM is an instance-based model. This implies that entities can never be grouped using the ‘OR’ operator. ← **this is not correct**

-

Future developments

Combining effect relations (NOTE: this is forward thinking, so it should not be part of the official spec)

Effect relations from more than one *activity unit* can be combined using logical operators:

- AND: The effect occurs only if all combined activities are executed.
- OR: The effect occurs if at least one of the combined activities is executed.

Contributors

V.0.1-0.3, October 2014 by Huaiyu Mi and Paul Thomas
V.0.7, June 21 2019 by Huaiyu Mi, Paul Thomas, Pascale Gaudet, Laurent-Philippe Albou, Benjamin Good.

Glossary

- Entity
- Occurrent
- Continuent
-