GO Annotation File (GAF) Format 2.1

Annotation data is submitted to the GO Consortium in the form of gene association files, or GAFs. This guide lays out the format specifications for GAF 2.1; for the previous GAF 2.0 file syntax, please see the GAF 2.0 file format guide.

For the first GAF 1.0 file syntax, please see the GAF 1.0 file format guide.

Please see the information on the changes in GAF 2.1.

Changes in GAF 2.1

GAF 2.1 **allows the use of pipes (|) and comma (,) in column 8 (with/from column)** compared to GAF 2.0 which allows the use of pipes only. **Pipe will indicate 'OR' and Comma will indicate 'AND'**.

In GAF 2.0, multiple values are separated by pipes where the pipe has been used to mean 'AND'. However, in the annotation extension field (column 16) pipe is used to indicate 'OR' and a comma to indicate 'AND'. This change to column 8 will allow consistent use of pipes and commas in the GO annotations. Please see the descriptions below for full details.

File Header

All gene association files must start with a single line denoting the file format, as follows:

```
!gaf-version: 2.1
```

Other information, such as contact details for the submitter or database group, useful link, etc., can be included in an association file by prefixing the line with an exclamation mark (`!`); such lines will be ignored by parsers.

Annotation File Fields

The annotation flat file format is comprised of 17 tab-delimited fields.

Column	Content	Required?	Cardinality	Example
1	DB	required	1	UniProtKB
2	DB Object ID	required	1	P12345
3	DB Object Symbol	required	1	PHO3
4	Qualifier	optional	0 or greater	NOT
5	GO ID	required	1	GO:0003993
6	DB:Reference (|DB:Reference)	required	1 or greater	PMID:2676709
7	Evidence Code	required	1	IMP
8	With (or) From	optional	0 or greater	GO:0000346
9	Aspect	required	1	F
10	DB Object Name	optional	0 or 1	Toll-like receptor 4
11	DB Object Synonym (|Synonym)	optional	0 or greater	hToll|Tollbooth
12	DB Object Type	required	1	protein
13	Taxon(|taxon)	required	1 or 2	taxon:9606
14	Date	required	1	20090118
15	Assigned By	required	1	SGD

16	Annotation Extension	optional	0 or greater	part_of(CL:0000576)
17	Gene Product Form ID	optional	0 or 1	UniProtKB:P12345-2

Definitions and requirements for field contents

DB (column 1)
refers to the database from which the identifier in **DB object ID** (column 2) is drawn. This is not necessarily the group submitting the file. If a UniProtKB ID is the **DB object ID** (column 2), DB (column 1) should be UniProtKB.
must be one of the values from the set of GO database cross-references (http://amigo.geneontology.org/xrefs)
this field is mandatory, cardinality 1

DB Object ID (column 2)
a unique identifier from the database in DB (column 1) for the item being annotated
this field is mandatory, cardinality 1
In GAF 2.1 format, the identifier **must reference a top-level primary gene or gene product identifier**: either a gene, or a protein that has a 1:1 correspondence to a gene. Identifiers referring to particular protein isoforms or post-translationally cleaved or modified proteins are *not* legal values in this field.
The **DB object ID** (column 2) is the identifier for the database object, which may or may not correspond exactly to what is described in a paper. For example, a paper describing a protein may support annotations to the gene encoding the protein (gene ID in **DB object ID** field) or annotations to a protein object (protein ID in DB object ID field).

DB Object Symbol (column 3)
a (unique and valid) symbol to which **DB object ID** is matched
can use ORF name for otherwise unnamed gene or protein
if gene products are annotated, can use gene product symbol if available, or many gene product annotation entries can share a gene symbol this field is mandatory, cardinality 1
The **DB Object Symbol** field should be a symbol that means something to a biologist wherever possible (a gene symbol, for example). It is not an ID or an accession number (**DB object ID** [column 2] provides the unique identifier), although IDs can be used as a **DB object symbol** if there is no more biologically meaningful symbol available (e.g., when an unnamed gene is annotated).

Qualifier (column 4)
flags that modify the interpretation of an annotation
one (or more) of `NOT` , contributes_to, colocalizes_with
this field is not mandatory; cardinality 0, 1, >1; for cardinality >1 use a pipe to separate entries (e.g. `NOT` |contributes_to)
See also the documentation on qualifiers (http://geneontology.org/GO.annotation.conventions.shtml#qual) in the GO annotation guide

GO ID (column 5)
the GO identifier for the term attributed to the **DB object ID**
this field is mandatory, cardinality 1

DB:Reference (column 6)

one or more unique identifiers for a single source cited as an authority for the attribution of the GO ID to the **DB object ID**. This may be a literature reference or a database record. The syntax is DB:accession_number.
Note that **only one reference can be cited on a single line** in the gene association file. If a reference has identifiers in more than one database, multiple identifiers for that reference can be included on a single line. For example, if the reference is a published paper that has a PubMed ID, we strongly recommend that the PubMed ID be included, as well as an identifier within a model organism database. Note that if the model organism database has an identifier for the reference, that identifier should **always** be included, even if a PubMed ID is also used.
this field is mandatory, cardinality 1, >1; for cardinality >1 use a pipe to separate entries (e.g. SGD_REF:S000047763|PMID:2676709).

Evidence Code (column 7)
see the GO evidence code guide (/book/guide-go-evidence-codes/) for the list of valid evidence codes for GO annotations
this field is mandatory, cardinality 1

With [or] From (column 8)
Also referred to as **with, from** or the **with/from** column
some examples are:

- DB:gene_symbol
- DB:gene_symbol[allele_symbol]
- DB:gene_id
- DB:protein_name
- DB:sequence_id
- GO:GO_id
- CHEBI:CHEBI_id
- IntAct:Complex _id
- RNAcentral:RNAcentral_id
- more...

This field is used to hold an additional identifier for annotations, for example, it can identify another gene product to which the annotated gene product is similar (ISS) or interacts with (IPI). An entry in the With/From field is not allowed for annotations made using the following evidence codes; EXP, IDA, IEP, TAS, NAS, ND. However, population of the With/From is mandatory for certain evidence codes, see the documentation for the individual evidence codes for more information. Cardinality = 0 is not allowed for ISS annotations made after October 1, 2006.
Multiple entries are allowed in the With/From field of certain evidence codes (see below) and they must be separated with a pipe or a comma. The pipe (|) specifies an independent statement (OR) and is equivalent to making separate annotations, i.e. not all conditions are required to infer the annotated GO term. The comma (,) specifies a connected statement (AND) and indicates that all conditions are required to infer the annotated GO term. In this case, 'OR' is a weaker statement than 'AND', therefore will be correct in all cases. Pipe and comma separators may be used together in the same With/From field.
This field is not mandatory overall, but is required for some evidence codes (see below and the evidence code documentation (/book/guide-go-evidence-codes/) for details); cardinality 0, 1, >1; for cardinality >1 use a pipe or comma to separate entries depending on the data as shown in the examples below.
The With/From field may be populated with multiple identifiers when making annotations using the following evidence codes: IMP, IGI, IPI, IC, ISS, ISA, ISO, ISM, IGC, IBA, IKR, RCA, IPI, IEA.
Annotations made using the following evidence codes may only use the pipe operator in the With/From field: ISS, ISA, ISO, ISM, IBA, IKR, RCA, IEA. It is not mandatory to use pipes, however, and some groups may prefer to make separate annotations.

Annotations made using the following evidence codes may use the pipe or comma operators in the With/From field: IPI, IMP, IGI, IC, IGC.
The **with** column may **not** be used with the evidence codes IDA, TAS, NAS, or ND.
Examples

1. Recording gene IDs for allelic variations in the With/from column for IMP evidence code: Multiple pipe-separated values in the with/from field indicate that the process is inferred from each perturbation independently. If more than one variation within the same locus resulted in a phenotype, those variations should be comma-separated (implying AND).
   For e.g. Two different deletion mutations and one RNAi inactivation support the same GO annotation for a Worm gene. The alleles are Pipe-separated in the With/From for this annotation: WB:WBVariation00091989|WB:WBVar00249869|WB:WBRNAi00084583
2. Recording gene IDs for mutants in the With/from column for IGI evidence code: Pipe-separated (OR) values should be used to indicate individual genetic interactions that result in the same inference for a process. Multiple values indicating triple mutants, for example, should be comma-separated (AND).
   For e.g. A triple mutant in C. elegans supports annotation to a specific process using IGI evidence. The gene identifiers are comma-separated in the With/From for this annotation indicating that the process is inferred from all three genes together: WBGene00000035,WBGene00000036
3. Recording IDs in the With/From column for IEA evidence code: Multiple, pipe-separated InterPro accessions are used for IEA-based annotations in the UniProt files and indicate individual (unconnected) inferences.
   For e.g. annotations to cell redox homeostasis (GO:0045454) that are inferred from three InterPro domains: InterPro:IPR005746|InterPro:IPR013766|InterPro:IPR017937

   This removes a large amount of redundancy and significantly decreases the size of UniProt files.

Note that a gene ID may be used in the **with** column for a IPI annotation, or for an ISS annotation based on amino acid sequence or protein structure similarity, if the database does not have identifiers for individual gene products. A gene ID may also be used if the cited reference provides enough information to determine which gene ID should be used, but not enough to establish which protein ID is correct.
'GO:GO_id' is used only when the evidence code is IC, and refers to the GO term(s) used as the basis of a curator inference. In these cases the entry in the 'DB:Reference' column will be that used to assign the GO term(s) from which the inference is made. This field is mandatory for evidence code IC.
The ID used in the with/from field should be an identifier for an individual entry in a database (such as a sequence ID, gene ID, GO ID, etc.). Identifiers from the Center for Biological Sequence Analysis (CBS), however, represent tools used to find homology or sequence similarity; these identifiers should not be used in the **with** column.

Aspect (column 9)
refers to the namespace or ontology to which the **GO ID** (column 5) belongs; one of P (biological process), F (molecular function) or C (cellular component)
this field is mandatory; cardinality 1

DB Object Name (column 10)
name of gene or gene product
this field is not mandatory, cardinality 0, 1 [white space allowed]

DB Object Synonym (column 11)
Gene symbol [or other text] Note that we strongly recommend that gene synonyms are included in the gene association file, as this aids the searching of GO.

this field is not mandatory, cardinality 0, 1, >1 [white space allowed]; for cardinality >1 use a pipe to separate entries (e.g. YFL039C|ABY1|END7|actin gene)

DB Object Type (column 12)
A description of the type of gene product being annotated. If a **gene product form ID** (column 17) is supplied, the **DB object type** will refer to that entity; if no **gene product form ID** is present, it will refer to the entity that the **DB object symbol** (column 2) is believed to produce and which actively carries out the function or localization described. one of the following: protein_complex; protein; transcript; ncRNA; rRNA; tRNA; snRNA; snoRNA; any subtype of ncRNA in the Sequence Ontology. If the precise product type is unknown, gene_product should be used. this field is mandatory, cardinality 1
The object type (gene_product, transcript, protein, protein_complex, etc.) listed in the **DB object type** field **must** match the database entry identified by the **gene product form ID**, or, if this is absent, the expected product of the **DB object ID**. Note that **DB object type** refers to the database entry (i.e. it represents a protein, functional RNA, etc.); this column does not reflect anything about the GO term or the evidence on which the annotation is based. For example, if your database entry represents a protein-encoding gene, then protein goes in the **DB object type** column. The text entered in the **DB object name** and **DB object symbol** should refer to the entity in **DB object ID**. For example, several alternative transcripts from one gene may be annotated separately, each with the same gene ID in DB object ID, and specific gene product identifiers in **gene product form ID**, but list the same gene symbol in the **DB object symbol** column.

Taxon (column 13)
taxonomic identifier(s) For cardinality 1, the ID of the species encoding the gene product. For cardinality 2, to be used only in conjunction with terms that have the biological process term multi-organism process or the cellular component term host cell as an ancestor. The first taxon ID should be that of the organism encoding the gene or gene product, and the taxon ID after the pipe should be that of the other organism in the interaction. this field is mandatory, cardinality 1, 2; for cardinality 2 use a pipe to separate entries (e.g. taxon:1|taxon:1000) See the GO annotation conventions for more information on multi-organism terms.

Date (column 14)
Date on which the annotation was made; format is YYYYMMDD
this field is mandatory, cardinality 1

Assigned By (column 15)
The database which made the annotation
one of the values from the set of GO database cross-references (http://geneontology.org/cgi-bin/xrefs.cgi)
Used for tracking the source of an individual annotation. Default value is value entered as the DB (column 1).
Value will differ from column 1 for any annotation that is made by one database and incorporated into another.
this field is mandatory, cardinality 1

Annotation Extension (column 16)
one of:

- DB:gene_id
- DB:sequence_id
- CHEBI:CHEBI_id
- Cell Type Ontology:CL_id
- GO:GO_id

Contains cross references to other ontologies that can be used to qualify or enhance the annotation. The cross-reference is prefaced by an appropriate GO relationship; references to multiple ontologies can be entered. For example, if a gene product is localized to the mitochondria of lymphocytes, the GO ID (column 5) would be mitochondrion ; GO:0005439, and the **annotation extension** column would contain a cross-reference to the term lymphocyte from the Cell Type Ontology (http://www.obofoundry.org/cgi-bin/detail.cgi?id=cell).
Targets of certain processes or functions can also be included in this field to indicate the gene, gene product, or chemical involved; for example, if a gene product is annotated to protein kinase activity, the annotation extension column would contain the UniProtKB protein ID for the protein phosphorylated in the reaction.
See the documentation on using the annotation extension column (http://geneontology.org/GO.annotation.extension.shtml) for details of practical usage; a wider discussion of the annotation extension column can be found on the GO wiki (http://wiki.geneontology.org/index.php/Annotation_Extension).
this field is optional, cardinality 0 or greater

Gene Product Form ID (column 17)
As the DB Object ID (column 2) entry *must* be a canonical entity—a gene OR an abstract protein that has a 1:1 correspondence to a gene—this field allows the annotation of specific variants of that gene or gene product. Contents will frequently include protein sequence identifiers: for example, identifiers that specify distinct proteins produced by to differential splicing, alternative translational starts, post-translational cleavage or post-translational modification. Identifiers for functional RNAs can also be included in this column.
The identifier used must be a standard 2-part global identifier, e.g. UniProtKB:OK0206-2

- When the **gene product form ID** (column 17) is filled with a **protein** identifier, the value in **DB object type** (column 12) must be **protein**. Protein identifiers can include UniProtKB (http://www.uniprot.org/help/uniprotkb) accession numbers, NCBI NP (http://www.ncbi.nlm.nih.gov/protein) identifiers or Protein Ontology (PRO) (http://pir.georgetown.edu/pro/pro.shtml) identifiers.
- When the gene product form ID (column 17) is filled with a **functional RNA** identifier, the **DB object type** (column 12) must be either **ncRNA**, **rRNA**, **tRNA**, **snRNA**, or **snoRNA**.

This column may be left blank; if so, the value in **DB object type** (column 12) will provide a description of the expected gene product.
More information and examples are available from the GO wiki page on column 17 (http://wiki.geneontology.org/index.php/GAF_Col17_GeneProducts).

Note that several fields contain database cross-reference (dbxrefs) in the format dbname:dbaccession. The fields are: **GO ID** [column 5], where dbname is always GO; **DB:Reference** (column 6); **With or From** (column 8); and **Taxon** (column 13), where dbname is always taxon. For GO IDs, do not repeat the 'GO:' prefix (i.e. always use GO:0000000, not GO:GO:0000000)

Helpdesk (/form/contact-go) • Citation/attribution (/page/go-citation-policy) • Terms of use (/page/use-and-license)

Member of the Open Biological and Biomedical Ontologies (http://www.obofoundry.org/)

The Gene Ontology Consortium is supported by a P41 grant from the National Human Genome Research Institute (NHGRI) [grant 2U41HG002273-17 (https://projectreporter.nih.gov/project_info_details.cfm?aid=9209989)].