Gene Curation Tool Database
Design Document
Version: 3.1
Date: July 21, 2015

---

1. Introduction

The philosophy that is ontology can be dated as far back as 1613; its practice however can be dated as far back as to Aristotle [11,12]. Philosophical ontology has "sought the definitive and exhaustive classification of entities in all spheres of being" [11]. Since the scientific revolution, the idea of ontology has been adopted by various domains (professions) one of the prominent being computer science [11]. Computer science uses ontologies to form relationships between concepts and by doing so it allows them to see the ontologies semantic structure viz. to see what it looks like. In addition, it allows for computer scientist to efficiently search and run "computational reasoning" on these structures [13]. Alternatively, ontologies can be classified into two major types: "domain ontologies" and "upper-level ontologies" [10]. Domain and upper-level ontologies, represent a vocabulary (concepts) for a domain of knowledge and describes general knowledge that carries across various disciplines [10]. Not to mention by adding annotations (information about a gene) to these concepts it allows for us to make connections between things that we may not have seen otherwise. By using ontologies and annotation, one can make collaboration easier, because we are able to share "meaning" using their structure [10]. In the following paragraphs, one will be acquainted with various types of ontologies that are used the domain of biology and why is visualizing these domain concepts as an ontology is crucial.

One of the most vital ontologies in biology is gene ontologies, which is often abbreviated GO. Gene Ontology essentially sets out to create a unified vocabulary for genes and to use the relationship between these concepts to see how species relate. For example, one could use this ontology to conclude that a specific protein found in eukaryotic cells relates to "core biological process" and is common among all eukaryotic cells [9]. Furthermore, one can see that understanding how genes among various species and how they relate can be vital when trying to understand the prolific amount of species that are around today.

Another vital type of ontology is the plant ontology, commonly abbreviated PO. The plant ontology is "a structured vocabulary and database resource that links plant anatomy and development to gene expression and phenotypic datasets from all areas of plant ontology" [16]. Just like other ontologies it uses relationships (e.g. is_a and part_of are the most common) to interpret how things relate [16]. Furthermore, there are various other ontologies that relate to plants. For example, the trait ontology (abbreviated TO). The trait ontology basic idea is that one can represent a trait as its own unique feature or characteristic. For example, plant height is

a type of vocabulary that is associated with the trait ontology [15]. These are just to name a few ontologies, so keep in mind there are many other ontologies that encompasses many other domains.

In conclusion, ontologies helps us understand the meaning of things and how they relate. By assisting one with understanding meaning, it helps speed up the discovery process. In any case, one can conclude that with the rate of discoveries increases so will innovation and innovation is the backbone of any profession.

1.1. Purpose

With the human population growing at such a rapid rate there is also a growing need to research plants, because they are the primary food source for many organisms on Earth [2]. Our hope is to speed up the plant research process by combining all the various plant ontology groups into one source. This will allow researchers who have various domains of knowledge to collaborate with one another. In addition, users will be able to easily search annotations that are associated with a gene, an ontology, etc. In conclusion, by giving researchers an easier way to collaborate with one another and by giving them an easy way to search through and manage annotations, we will speed up the overall research process.

1.2. Scope

The scope of this project is to design a database in that will merge the many branches of ontology. We will be creating a collaborative web portal that will interact with our database and allow users to easily share results (e.g. annotation data). This web portal will act as a wiki so to speak, in that it allows the users to easily manage annotations. Furthermore, we will implement our database in such a way that users can easily search annotations associated with a gene, an ontology, etc. In addition, we will be making API calls to SOLR, which is the data store AmiGo is using, to query various information for the many ontologies; this is to ensure that we always have the most up-to-date version. Lastly, our web portal will allow us to easily bulk upload our approved annotation to AmiGo, to ensure that they always have our most up-to-date annotation information.

1.3. Terminology / Definitions

- **Annotations** - information about a gene that is attached to these vocabularies (concepts) in ontologies and used to describe their relationships. Often contain an evidence code and literature associated with it to back up this newly found information about a gene according to Dr. Pankaj.

- **AmiGO** - A web tool for accessing the Gene Ontology project's data (including browsing genes and their corresponding annotations).
- **cROP** - "Common Reference Ontology for Plants", a set of ontologies concerning plants [3]. This was the name of the project before it was changed to Planteome by Dr. Pankaj.
- **Gene Ontology** - The Gene Ontology refer to vocabulary applied to all gene and protein roles in cells. Which including three main parts: the biological process (p), molecular function (f) and cellular component (c). [7]
- **Genomes - The complete genetic material of an organism [5].**
- **MapReduce - is a programming model that where you split up data across processes so that they can be ran independently in parallel [2].**
- **NoSQL - "Not Only SQL", steer away from your traditional relational database model for better performance on flat data [2].**
- **Ontologies - Ontologies have long been used in an attempt to describe all entities within an area of reality and all relationships between those entities. An ontology comprises a set of well-defined terms with well-defined relationships. The structure itself reflects the current representation of biological knowledge as well as serving as a guide for organizing new data. Data can be annotated to varying levels depending on the amount and completeness of available information. This flexibility also allows users to narrow or widen the focus of queries. Ultimately, an ontology can be a vital tool enabling researchers to turn data into knowledge [7].**
- **Phenotypes - observational characteristics of an organism [5].**
- **Protein - "are large biomolecules, or macromolecules, consisting of one or more long chains of amino acid residues" [8].**

1.4. Overview

The rest of this document will discuss the specifics the the Planteome software project. First, we will be providing a much less technical design overview, so that users can get a feel of how the Planteome software will work. Lastly, we will be giving a more thorough idea of the design process that will aid developers on developing the software.

1.5. References/Suggested Readings

#	Topic	Reference
1	iPlant	http://www.iplantcollaborative.org/about-iplant
2	Planteome Project Proposal	http://wiki.planteome.org/images/1/1d/Planteome_grant_NSF_-1340112.pdf
3	G8 Planteome Presentation Slides	G8_open_data_2013_DC.pdf

4	*Genome* definition	https://en.wikipedia.org/wiki/Genome
5	*Phenotype* definition	https://en.wikipedia.org/wiki/Phenotype
6	Gene Ontology	http://biochem218.stanford.edu/Projects%202010/Blair%202010.pdf
7	*Gene Ontology* Definition	http://www.ncbi.nlm.nih.gov/pubmed/10802651
8	*Protein* definition	https://en.wikipedia.org/wiki/Protein
9	Gene Ontology: tool for unification of biology	http://www.nature.com/ng/journal/v25/n1/full/ng0500_25.html
10	Some concepts on ontologies, semantic nets, and taxonomy	http://ascelibrary.org/doi/pdf/10.1061/(ASCE)0887-3801(2005)19:4(394)
11	History of Ontology Info	http://www.cs.vassar.edu/~weltyc/papers/fois-intro.pdf
12	More History on Ontology	http://scholar.google.com/scholar_url?url=http://www.aaai.org/ojs/index.php/aimagazine/article/download/1714/1612&hl=en&sa=X&scisig=AAGBfm1z8Q2ToahURBrCayYXbQ7uhkuf1g&nossl=1&oi=scholarr
13	Gene Ontology Annotations	http://nar.oxfordjournals.org/content/37/suppl_1/D555.full.pdf+html
14	Ontologies in Relational Databases	http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.2534&rep=rep1&type=pdf
15	TO Information	http://onlinelibrary.wiley.com/doi/10.1002/cfg.156/epdf
16	Information on Plant Ontology	http://www.researchgate.net/publication/232271560_An_extension_of_the_Plant_Ontology_project_supporting_wood_anatomy_and_development_research

2. Design Overview

2.1. Features

This web portal will have a plethora of features. Here is a list of the various features:

- Relational Database System
    - We plan to use a relational database because in this stage of development our client (Planteome group) doesn't full know the type of queries they want to make and to take full advantage of NoSQL databases a user has to have a good idea of what type of queries they will be querying.
- Role Based Access Control (RBAC)
    - This will be used to distinguish the permissions of the various user levels and what they will have access to. It will also ensure that we have a well structured user level/role design which increases security.
- User Management System
    - We want the managers and admins to easily be able to manage the users (e.g. banning users)
- Users will be able to have specialties in species
    - This feature will be added to better assign annotations to approve/disapprove to managers. The idea is to give them annotations in areas they specialize in
- Users can can flag a annotation as valid or invalid
    - The idea behind this is once an annotation invalid/valid flag ratio hits a certain threshold a manager specializing in that annotation will be alerted to review this annotation.
- Efficient ways to search and manage annotations
    - This is a crucial feature to our system, because it is the core backbone of what we wanna do. We want users to be able to easily search and manage annotations.
- Save annotation draft for later
    - This is to ensure that if it is late in the night or they have to do something they can save it and finish it at a later date.
- Two types of comments on annotations
    - The two types of comments on annotations will be private and public comments. The private comments will be just for the approving manager to see and the public comment all users will be able to see on this annotation.

2.2. Basic User Level Hierarchy

Figure 1:

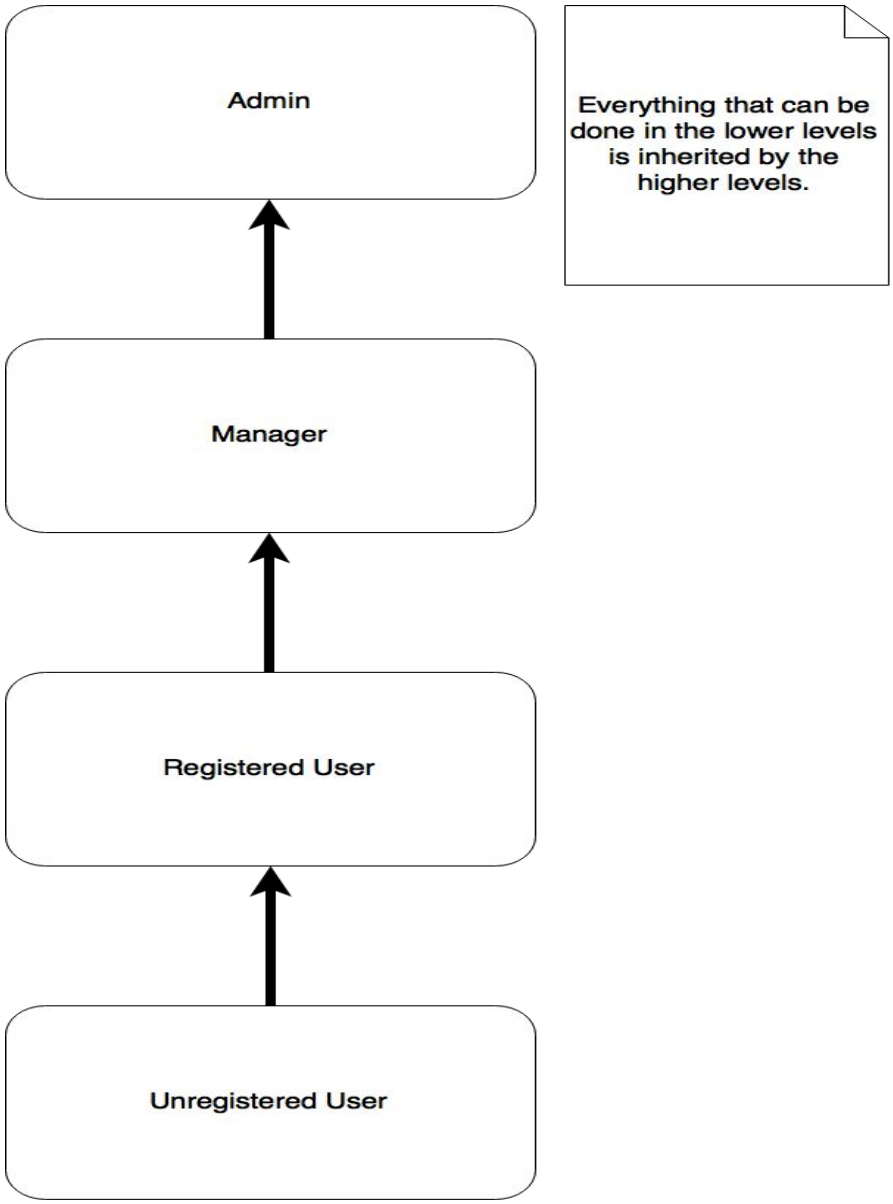

Figure 1: *This diagram essentially shows the various user levels a user can have and based on these user levels a user will have different permissions and access. The idea is that the lowest level user can be found at the bottom and as you go up in the hierarchy the user gains access to more permissions and roles and it also retains all the permissions and roles defined in previous layers. So this diagram serves as a gateway for the user to essentially understand the flow of user levels.*

2.3 Use Case Diagrams

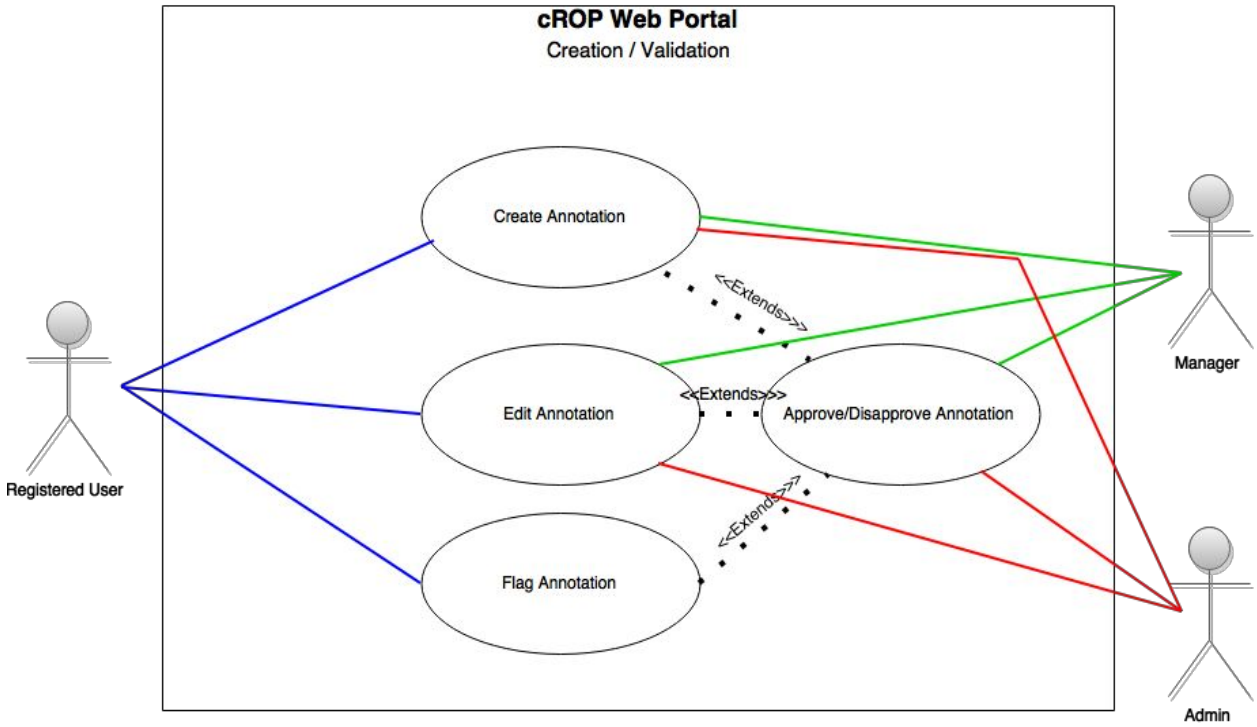

Figure 2: *This diagram shows the basic process of creating, editing, and flagging (e.g. All three of these actors can create, edit, and flag annotations, but only the managers and admin can approve/disapprove an annotation). Purple lines representing unregistered users, blue line representing registered users, green line representing managers, and the red line representing admins. One can also see that each time an annotation is edited, flagged (a certain number of times) and an annotation is created it will be prompted to be approved or disapproved (represented in the use case diagram by extend) by a manager or admin with a specialty in that area.*

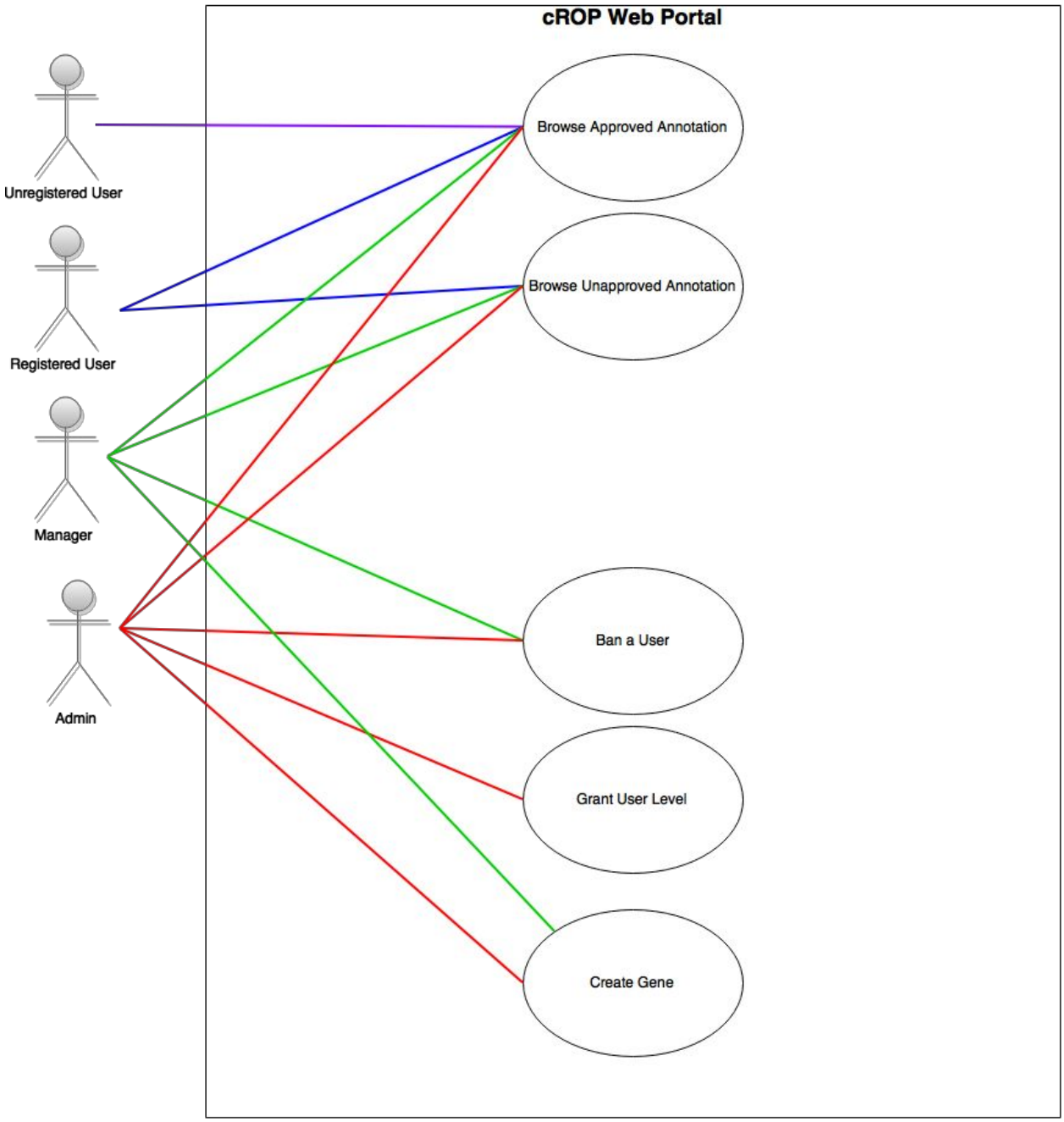

Figure 3: *This diagram shows the various actions that the various levels of users can do. Purple lines representing unregistered users, blue line representing registered users, green line representing managers, and the red line representing admins.*

2.4 ER Diagram (Database Design)

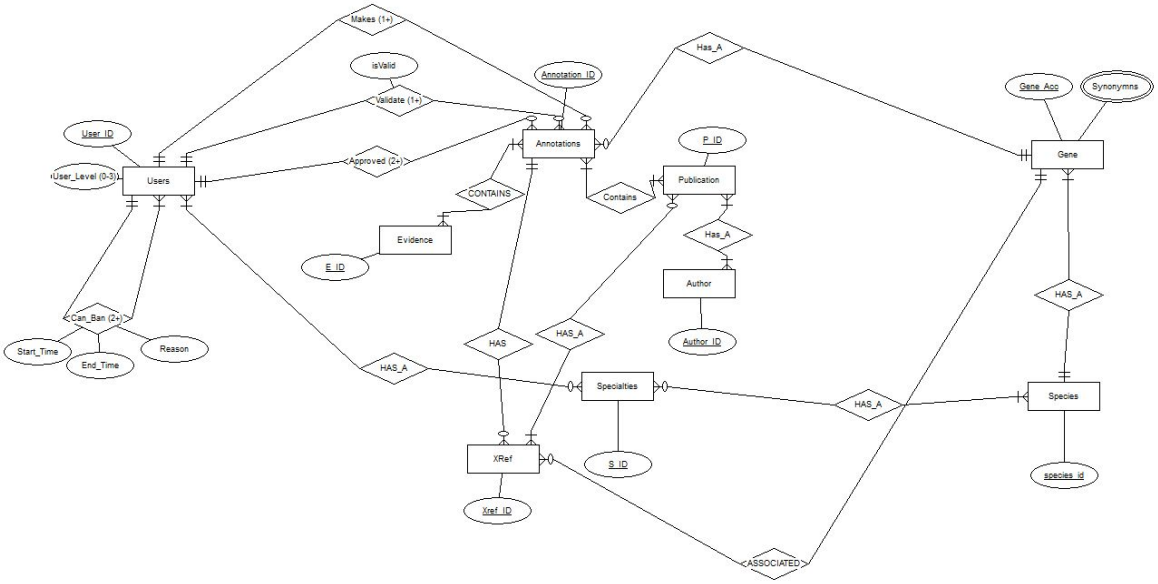

Figure 4: *This diagram is a basic ER diagram of our web portal. Where we show the relationships (diamond shaped) between entities (box shaped). One can read a relationship like this, the entity gene has to have one species and that species is mandatory. Another example could be that the entity gene has an many annotations or no annotations, so we call that optional to many.*

2.5 Table Structure Diagram

Users

ID	Unique ID
Username	Unique Username
Password	Password
email	user's email
Last_IP	The IP of last login
Last_Login_Timestamp	The time of last login
User_Level	Can be 0,1,2 or 3 (Defines the level of the user)

credit	the credit for the contribution of the system (maybe split it into its own table)

User_banned

banned_id	Unique ID
Manager_id	Manger who banned the user
Start_time	Start ban time
End_time	End ban time
Reason	Reason for banning

Specialty

id	Specialty ID
name	Specialty Name
Species_id	Species ID

User_Specialty

User_id	User ID
Specialty_id	Specialty ID

Gene

Gene_ACC	Unique Gene ID
Gene_name	Gene Name
Gene_description	Gene Description
Chromosome number	Genome_location
Start	Chromosome Start
End	Chromosome End

Synonyms

Gene_ACC	Unique Gene ID
Synonym_name	Synonym Name

Species

Species_id	Species ID
Species_name	Species Name

Gene_Species

Gene_ACC	Gene ID
Species_id	Species ID

Annotation

Index	Unique ID
Annotation_ID	Annotation ID
Ontology_ACC	Ontology ID
Gene_ACC	Gene ID
Created_by	Who added the annotation (user_ID)
Approved_by	Who approved (user_ID)
Created_date	Date submitted
Approved_date	Date Annotation Approved (Can be null)
Comment	every people can read comment
Note	Private and not readable for others
Draft_flag	0: in processing 1: submitted

Annotation_Validation

Annotation_ID	Annotation ID
User_ID	User ID
isValid	-1 invalid or 1 valid
date	Date
reason	Reason why flagged

Annotation_Approvement

Annotation_ID	Annotation ID
User_ID	User ID
isApprovement	-1 approvement or 1 approvement
date	Date
reason	Reason why

Approved_Annotations

Annotation_ID	Annotation ID
Ontology_ACC	Ontology ID
Gene_ACC	Gene ID
Created_by	Who added the annotation (user_ID)
Approved_by	Who approved (user_ID)
Created_date	Date submitted
Approved_date	Can be NULL who approved it
Comment	Everyone can read this comment
Note	Private and not readable for others

Evidence

Evidence_ID	Evidence ID
Evidence_Code	Evidence Code

Annotation_Evidence

Annotation_ID	Annotation ID
Evidence_ID	Evidence ID

Publications

title	Name of Publication
Publication_ID	Publication ID
Source	Source
Abstract	Abstract Text

Author

Author_ID	Author ID
Author_first_name	Author's First Name
Author_last_name	Author's Last Name

Author_Publication

Author_ID	Author ID
Publication_ID	Publication ID

Xref

XRef_name	XRef Name
Xref_ ID	Xref_Object_ID

Xreference_relation

XRef_ID	XRef ID
Object_ID	Object Link