**Gene Curation Tool (GCT)**

**Design Document**

08/24/2015

Contents

[1 Introduction 5](#_Toc428313797)

[1.1 Purpose 5](#_Toc428313798)

[1.2 Basic Genetic knowledge 5](#_Toc428313799)

[1.3 Ontology 7](#_Toc428313800)

[1.4 Annotation 9](#_Toc428313801)

[1.5 Planteome Project 9](#_Toc428313802)

[1.6 Gene Curation Tool (GCT) 10](#_Toc428313803)

[2 Related Ontology Databases 11](#_Toc428313804)

[2.1 Amigo and Amigo2 11](#_Toc428313805)

[2.1.1 Introduction 11](#_Toc428313806)

[2.1.2 Features 12](#_Toc428313807)

[2.1.3 GO Annotation File (GAF) Format 12](#_Toc428313808)

[2.2 Gramene 13](#_Toc428313809)

[2.3 AgriGO 13](#_Toc428313810)

[2.4 Conclusion 13](#_Toc428313811)

[3 Software Requirement Specification 13](#_Toc428313812)

[3.1 Product Perspective 13](#_Toc428313813)

[3.2 Scope 14](#_Toc428313814)

[3.3 Operating Environment 14](#_Toc428313815)

[3.4 Role Based Access Control (RBAC) 14](#_Toc428313816)

[3.5 Product Functions 14](#_Toc428313817)

[3.6 User management subsystem 16](#_Toc428313818)

[3.6.1 Use Case: Register 16](#_Toc428313819)

[3.6.2 Use Case: Login 18](#_Toc428313820)

[3.6.3 Use Case: Ban User 18](#_Toc428313821)

[3.6.4 Use Case: User role Hierarchy 19](#_Toc428313822)

[3.6.5 Use Case: Credit of the contribution 20](#_Toc428313823)

[3.6.6 Use Case: Edit Specialty 21](#_Toc428313824)

[3.7 Annotation management subsystem 21](#_Toc428313825)

[3.7.1 Use Case: browse/edit/add annotation 21](#_Toc428313826)

[3.7.2 Use Case: Save annotation draft 22](#_Toc428313827)

[3.7.3 Use Case: Save note 22](#_Toc428313828)

[3.7.4 Use Case: flag annotation 22](#_Toc428313829)

[3.7.5 Use Case: Comment on annotation 22](#_Toc428313830)

[3.8 Object management subsystem 23](#_Toc428313831)

[3.9 Publication management subsystem 24](#_Toc428313832)

[4 API design 25](#_Toc428313833)

[4.1 Object import API 25](#_Toc428313834)

[4.2 Annotation import API 25](#_Toc428313835)

[4.3 Annotation export API 25](#_Toc428313836)

[4.4 Utilize API to get Ontology information 25](#_Toc428313837)

[5 User Interface Design 26](#_Toc428313838)

[6 Database Design 26](#_Toc428313839)

[6.1 ER diagram 26](#_Toc428313840)

[6.2 Tables design 26](#_Toc428313841)

[6.2.1 Table: Users 26](#_Toc428313842)

[6.2.2 Table: User\_banned 27](#_Toc428313843)

[6.2.3 Table: Specialty 27](#_Toc428313844)

[6.2.4 Table: User\_Specialty 27](#_Toc428313845)

[6.2.5 Table: Object 28](#_Toc428313846)

[6.2.6 Table: Synonyms 28](#_Toc428313847)

[6.2.7 Table: Species 28](#_Toc428313848)

[6.2.8 Table: Gene\_Species 28](#_Toc428313849)

[6.2.9 Table: Annotation 29](#_Toc428313850)

[6.2.10 Table: Annotation\_Validation 29](#_Toc428313851)

[6.2.11 Table: Annotation\_Approvement 30](#_Toc428313852)

[6.2.12 Table: Approved\_Annotations 30](#_Toc428313853)

[6.2.13 Table: Evidence 31](#_Toc428313854)

[6.2.14 Table: Annotation\_Evidence 31](#_Toc428313855)

[6.2.15 Table: Publications 31](#_Toc428313856)

[6.2.16 Table: Author 31](#_Toc428313857)

[6.2.17 Table: Author\_Publication 32](#_Toc428313858)

[6.2.18 Table: Xref 32](#_Toc428313859)

[6.2.19 Table: Xreference\_relation 32](#_Toc428313860)

[7 References 32](#_Toc428313861)

[Appendix A: Definitions and Abbreviations 33](#_Toc428313862)

# Introduction

In this section, we will try to introduce some basic but necessary information and knowledge to the participant of this project.

## Purpose

This document is intended for people developing and/or supervising the Gene Curation Tool (GCT) project. We want to make a clarification of the related terminologies and the purpose of this project, and the requirements of the users for all of the attendants. So the beginners could keep the pace with the process as soon as possible, which is absolutely necessary for such a long term project. The whole document could be divided into two parts, the introductions and the design.

In first section and the second section, we try to introduce all necessary terms and background knowledge to help the beginner understand the big picture of our project and related biological terms. Also, the big picture of the project will be delineate to help the readers catch the main points. Besides, we will also introduce some similar system for comparison, so the readers could get the research status readily.

From the third section, we start to write down the details of design and implement as much as we could, so biologists, programmers and researchers could keep synchronous. And more necessary, in the future, when other participant need to modify or upgrade this project, they would not be confused by the codes and terms we are using now.

## Basic Genetic knowledge

Following terms are ordered by the inclusion relationship, from the smaller unit to bigger ones.

*Gene*

A gene is a locus (or region) of DNA that encodes a functional RNA or protein product, and is the molecular unit of heredity. The transmission of genes to an organism's offspring is the basis of the inheritance of phenotypic traits. Most biological traits are under the influence of polygenes (many different genes) as well as the gene–environment interactions.

DNA

Deoxyribonucleic acid (DNA) is a molecule that carries most of the genetic instructions used in the development, functioning and reproduction of all known living organisms and many viruses. DNA is a nucleic acid; alongside proteins and carbohydrates, nucleic acids compose the three major macromolecules essential for all known forms of life. Most DNA molecules consist of two biopolymer strands coiled around each other to form a double helix.

RNA

Ribonucleic acid (RNA) is a polymeric molecule implicated in various biological roles in coding, decoding, regulation, and expression of genes. RNA and DNA are nucleic acids, and, along with proteins and carbohydrates, constitute the three major macromolecules essential for all known forms of life. Like DNA, RNA is assembled as a chain of nucleotides, but unlike DNA it is more often found in nature as a single-strand folded onto itself, rather than a paired double-strand. Cellular organisms use messenger RNA (mRNA) to convey genetic information that directs synthesis of specific proteins. Many viruses encode their genetic information using an RNA genome.

Chromosome

A chromosome is a packaged and organized structure containing most of the DNA of a living organism. It is not usually found on its own, but rather is complexed with many structural proteins as well as associated transcription (copying of genetic sequences) factors and several other macromolecules. Two "sister" chromatids (half a chromosome) join together at a protein junction called a centromere.

*Genome*

In modern molecular biology and genetics, the genome is the genetic material of an organism. It consists of DNA (or RNA in RNA viruses). The genome includes both the genes and the non-coding sequences of the DNA/RNA

*Germplasm*

Germplasm is the living genetic resources such as seeds or tissue that is maintained for the purpose of animal and plant breeding, preservation, and other research uses. These resources may take the form of seed collections stored in seed banks, trees growing in nurseries, animal breeding lines maintained in animal breeding programs or gene banks, etc. Germplasm collections can range from collections of wild species to elite, domesticated breeding lines that have undergone extensive human selection.

*QTL*

A quantitative trait locus (QTL) is a section of DNA (the locus) that correlates with variation in a phenotype (the quantitative trait). The QTL typically is linked to, or contains, the genes that control that phenotype. QTLs are mapped by identifying which molecular markers correlate with an observed trait. This is often an early step in identifying and sequencing the actual genes that cause the trait variation.

*Gene product*

A gene product is the biochemical material, either RNA or protein, resulting from expression of a gene. A measurement of the amount of gene product is sometimes used to infer how active a gene is.

Phenotype

A phenotype is the composite of an organism's observable characteristics or traits, such as its morphology, development, biochemical, behavior, phenology, physiological properties, and products of behavior (such as a bird's nest).

## Ontology

Ontology is the [philosophical](https://en.wikipedia.org/wiki/Philosophy) study of the nature of [being](https://en.wikipedia.org/wiki/Being), [becoming](https://en.wikipedia.org/wiki/Becoming_(philosophy)), [existence](https://en.wikipedia.org/wiki/Existence), or [reality](https://en.wikipedia.org/wiki/Reality), as well as the basic [categories of being](https://en.wikipedia.org/wiki/Category_of_being) and their relations, can be dated as far back as 1613, its practice however can be dated as far back as to Aristotle. Philosophical ontology has sought the definitive and exhaustive classification of entities in all spheres of being.

Since the scientific revolution, the idea of ontology has been adopted by various domains (professions) one of the prominent being computer science. Computer science uses ontologies to form relationships between concepts and by doing so it allows them to see the ontologies semantic structure visualization to see what it looks like. In addition, it allows for computer scientist to efficiently search and run “computational reasoning” on these structures.

In biology, a key element of a semantic data integration framework is the ontology, a formal representation of a knowledge domain in which concepts, terms and classes are interrelated in a graph or network, with edges indicating relationships between these concepts. Data are attached to terms through metadata and tags to make them discoverable online while the relationships between terms are read by a computer algorithm to aggregate the terms, and thus answer biological questions (such as those mentioned above) to infer various associations.

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.

*Gene Ontology (GO)*

One of the most vital ontologies in biology is gene ontologies, which is often abbreviated GO. Genomic sequencing has made it clear that a large fraction of the genes specifying the core biological functions are shared by all eukaryotes. Knowledge of the biological role of such shared proteins in one organism can often be transferred to other organisms [1]. GO provides annotation on three independent ontologies: processes, function and cellular compartment.

For example, one could use GO to conclude that a specific protein found in eukaryotic cells relates to “core biological process” and is common among all eukaryotic cells [1]. 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.

*Plant Ontology (PO)*

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 [19]. PO describes plant anatomy, morphology and growth and development stages to annotate the source plant sample harvested/evaluated for gene expression and/or phenotype assays. Just like other ontologies it uses relationships (e.g. is\_a and part\_of are the most common) to interpret how things relate [19].

Plant Trait Ontology (TO)

The trait ontology basic idea is that one can represent a phenotypic trait as its own unique feature or characteristic. Each trait is a distinguishable feature, characteristic, quality or phenotypic feature of a developing or mature plant, or a plant part. Currently, the TO has been organized around eight main trait classes: quality, yield, stress, stature or vigor, anatomy and morphology, biochemical, growth and development, sterility and fertility. The TO contains ~1300 terms, allowing it to encompass a broad range of plant traits.

The TO is a ‘composite’ or ‘pre-composed’ ontology, meaning the terms are assembled in advance from simpler ‘building block’ concepts. For example, the trait fruit shape is an observable characteristic composed using the Plant Ontology term fruit (PO: 0009001) and the PATO attribute (quality) shape (PATO: 0000052). In a species specific phenotype assay this trait term can be associated with different phenotype values. For example round (PATO: 0000411), to describe a fruit that has a round shape.

Plant Environment Ontology (EO)

The EO describes various types of experimental treatments a plant sample receives, and as described in the phenotype and gene expression assays. Currently, the structure of the EO consists of three main upper level classes: the biotic environment, abiotic environment and study types. The study types can be used to identify the growth study facility, such as field study, growth chamber, and greenhouse.

Plant Stress Ontology (PSO)

The PSO will describe both major types of stress: abiotic (drought, salinity, temperature, nitrogen deficiencies, etc.) and biotic (pests, pathogens, symbiotic organisms, competition, diseases, etc.), which will form the two main branches of the PSO.

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.

## Annotation

Annotation is the process of assigning ontology terms to gene or genetic resources such as germplasm or QTL. For example, a PhD student in 1990 might routinely sequence 1 kilobase (a unit of measurement in molecular biology) of DNA using some sequencing technology (subject to possessing technical skills to do the experiment). All these finding genes, sequencing parts of the genome and functional analysis is a process of annotation.

To clarify this definition, we still use the GO as an example. As Gene products have common functions across species, there are many proteins in more than 20 species are annotated to the GO term “*transcription factor binding”* in Uniprot (a database of proteins). Transcription factor binding is a high level term, it has 9 direct child terms, each of which also has child terms linked by is a relationships. The structure of the molecular hierarchy of the GO allows subsumption queries, which traverse these relationships and representing more or less specific knowledge about biological processes as the hierarchy is traversed. The GO enforces an “all paths to root must be true” rule so that the terms and their relationships represent a statement of biological truth based on available knowledge. For example. Transcription factor binding is-a protein binding is-a binding is-a molecular function.

More or less specific annotation can be specified by a curator selecting a term from lower or higher up the hierarchy. Annotations are made to GO based on scientific literature, automated analyses based on sequence homology and assertions made by expert curators. Annotations change over time on the basis of emerging biological knowledge, and the content of the GO also changes as terms are added, or removed, annotations are therefore updated periodically. In the example above we considered gene specific annotation and explored the use of the GO in the context of protein databases.

## Planteome Project

The Plant Genome Research Program (National Science Foundation) award funding for three years to The Planteome Project ([http://www.planteome.org](http://www.planteome.org/)), an international collaboration to support the development of “Common Reference Ontologies and Applications for Plant Biology” (cROP).

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 [4]. 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.

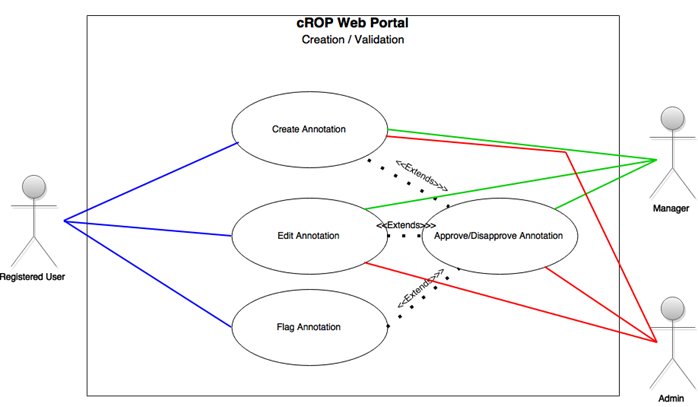
In conclusion, by giving researchers an easier way to collaborate with one another and by giving them an easy way to search through and manage annotations, we will speed up the overall research process. The goals of the Planteome Project are to provide researchers and agricultural breeding programs a common semantic framework, and a focused set of comparative analysis tools to leverage the scientific value of the ever-expanding array of sequenced plant genomes and phenotype data. The Planteome Project will develop a set of common data standards and universal reference vocabularies to describe plant biology and plant stresses, and standardized plant gene and phenotype annotation workflows.

The proposed project seeks to create a centralized platform where reference ontologies for plants will be used to access cutting-edge data resources for plant traits, phenotypes, diseases, genomes and semantically-queried genetic diversity and gene expression data across a wide range of plant species. The cROP will develop the Plant Trait Ontology (TO), the Plant Stress Ontology (PSO), and the Plant Environment Ontology (EO) besides taking over the development of Plant Ontology (PO). It will also include relevant aspects of ontologies such as Gene Ontology (GO), Cell type (CL), Chemical Entities (ChEBI), Protein Ontology and the Phenotypic Qualities Ontology (PATO). The cROP fits into the existing biological ontology landscape, and will be an active participant in the OBO Foundry, adhering to all of its principles. We will collaborate with the existing reference ontologies and contribute to their enrichment in terms and definitions of common importance to plant biology.

## Gene Curation Tool (GCT)

The development of Gene Curation Tool is one main aim of the Planteome project. Develop a community-wide standardized workflow and tools for ontology development, curation and improved annotation of genes, genomes, phenotype and germplasm. In order to maintain high quality standards for data annotation and provide a common place to find these annotations, we are proposing a new web portal and a data warehouse to host and serve the reference ontologies and the associated annotation data.

So the main jobs of the development of the Gene Curation Tool could be partitioned into two main parts, First, we need to develop a database which could save all these annotation information, object data, etc. Secondly, develop a website which could visit the database and efficiently show the data to the users, and users could easily handle the interaction between the data and web pages.



**Figure 1:** 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.

# Related Ontology Databases

In this section, we will delineate some ontology databases and their corresponding browse platform being used, present the basic information about them and address the strength and weakness of these Databases, through comparing these features, thus increase the capabilities and the competence while building the GCT.

## Amigo and Amigo2

Amigo has a deep relationship with our project. So in this section, we will demonstrate what is Amigo and how we will use utilize it in our project.

### Introduction

### Features

Build based on Gene Ontololgy

Justin is extending it to other ontology data (Amigo2)

Provide all ontology data

GCT is a backup and database could be easily modified

GCT would also how the data in process, unapproved, out date

### GO Annotation File (GAF) Format

Annotation File Fields (GAF 2.1)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 | 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 | optional | 0 or greater | hToll|Tollbooth |
| 12 | DB Object Type | required | 1 | protein |
| 13 | 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 |

## Gramene

Introduction:

Features:

## AgriGO

Introduction:

Features:

## Conclusion

As above statement, we conclude following features which would be provide great convenient for the biologists to use GCT:

# Software Requirement Specification

In this section, we will try to itemize the user requirements, which serve as a  
guide to the developers on one hand and a software validation document for the  
prospective client on the other.

## Product Perspective

GCT is aimed toward biologists come from all over the world who has considerable number of research achievement related to the genes and their functions, or who need to get the most affluent annotation information.

With using GCT, the scientists could easily browse the newest ontological annotation data, and also could be able to upload their own annotation data, edit the data from other scientists. They could also easily to communicate with other scientist by making comment on the data or flagging the annotation to valid or invalid.

Develop a GCT means create a community-wide standardized workflow for plant genome annotation. No matter what the species and ontology type the scientist specialized in, it would become much easier for them to share their annotation data and get other scientists research result.

## Scope

We descript the features in scope of GCT.

1. Design a database will merge the many branches of ontology.
2. Users could easily search and browse annotations associated with a gene, an ontology, etc.
3. Users could easily search and browse information for genes and ontologies, publications, authors, etc.
4. This web portal will act as a wiki so to speak. Biologists could easily share their research results. Which means they could upload their annotation easily. When they found the existed annotation is problematic, they could make suggestion or edit it as they want.
5. GCT will also works similar to a forum. That is people could easily comment on the existed annotations and communicate with others.
6. All the knowledge will be managed strictly by the admin of the system, every modification or creation of the knowledge will need be approved by admins before being shown to users.

## Operating Environment

Database: MySQL

\*We plan to use a relational database because in this stage of development, our client (Planteome group) doesn't fully 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.

Server: Apache

## 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.

## Product Functions

GCT should support the following use cases:

|  |  |  |  |
| --- | --- | --- | --- |
| Class of use cases | RS\_ID | Use cases | description |
| unregistered user’s capabilities (basic contributor) | 1-1 | register |  |
| 1-2 | browse annotation information |  |
| 1-3 | browse ontology information |  |
| 1-4 | browse gene information |  |
| 1-5 | comment on annotations |  |
| registered user’s capabilities (expert contributor) |  | Login |  |
|  | profile management  (include change password) |  |
| 1-2,1-3,1-4 | browse annotation/ontology/gene information |  |
|  | export annotation information |  |
|  | browse credit |  |
|  | edit annotation |  |
|  | save annotation draft |  |
|  | save note to user self |  |
|  | flag annotation |  |
|  | add annotation |  |
|  | comment on annotation |  |
| Admin’s capabilities |  | all capabilities of registered user |  |
|  | manage user’s information  (credit, password, profile) |  |
|  | manage users’ role |  |
|  | ban user from activities |  |
|  | approve annotation modifications  (edit, flag and add) |  |
|  | edit publication |  |
|  | edit evidence |  |
| Super admin’s capabilities |  | All capabilities of admin |  |
|  | Import data (gene data, annotation data, etc.) |  |

## User management subsystem

In this section, we will introduce all use cases related to the user’s profile information.

### Use Case: Register

Unregistered user could only browse the information from the web, every users of could register and then login to get higher access of the system.

|  |  |  |
| --- | --- | --- |
| **USE CASE: Register** | | |
| **Description** | User could register to become an expert contributor | |
| **Main Actor** | non-registered user | |
| **Trigger** | click the register button | |
| **Typical case Scenario** | **Action** | **Response** |
| 1 fill all required information |  |
| 2 click submit button | 3 system will check the information been filled |
|  | 4 system will save the data |
| **Alternate Scenario** | 4: if 3 found the information provided is not correct, the system will prompt a dialogue to indicate the problem | |
| **Result** | successfully create a new user | |
| **Constraint** | the user’s name should be an email address  the user need to select specialty from a drop list  the user name should be unique  the password need to be input twice, and both of them should be same | |

|  |  |  |
| --- | --- | --- |
| **INPUT** | | |
| **NAME** | | **DESCRIPTION** |
| name | first name | User’s name will be shown on the pages. |
| last name |
| middle name |
| affiliation | institute |  |
| XXX |  |
| Specialty | | the user could only edit the annotation belong to specific specialty |
| user\_name | | Email address, used to login, need to check if there is exist a same user\_name and the format of the user\_name is correct. |
| phone | | contact information |
| country | |  |
| password | | need to be input twice to confirm |

|  |  |
| --- | --- |
| **OUTPUT** | |
| **NAME** | **DESCRIPTION** |
| success |  |
| user\_name occupied |  |
| miss required information |  |

### Use Case: Login

The registered user could login to the system to get more abilities such as add, flag, edit the annotations and manage his own personal information.

|  |  |
| --- | --- |
| **INPUT: login** | |
| **NAME** | **DESCRIPTION** |
| user name | email address |
| password | need to be input twice to confirm |

|  |  |
| --- | --- |
| **OUTPUT: logins** | |
| **NAME** | **DESCRIPTION** |
| success |  |
| wrong password |  |
| no username |  |

### Use Case: Ban User

The managers should be able to ban the registered users from making any actions such as add annotations, suggesting the existed annotations to invalid, etc., for the security and in case of some irresponsible behaviors.

|  |  |
| --- | --- |
| **INPUT** | |
| **NAME** | **DESCRIPTION** |
| user\_name | the user need to be banned |
| duration | the persist time in which the user could not make any actions such as comment, edit annotation and make suggestion. |
| reason | why the user has been banned |

### Use Case: User role Hierarchy

The users are divided into four levels (0-3), different level correspond to different ability range. The admin (both super admin and admin) are able to manage the users’ roles. Also, the admins could be able to edit the personal information of the user in backstage.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Role | comment | edit | add | flag | approve | import |
| super admin | X | X | X | X | X | X |
| admin | X | X | X | X | X |  |
| contributor expert | X | X | X | X |  |  |
| basic contributor | X |  |  |  |  |  |

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.

**

**Figure 2:** 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.

### Use Case: Credit of the contribution

Each contribution of the annotation may lead to an accumulation of credit. The users could see the ranking result and get to know who contribute more to the whole gene curation system.

|  |  |  |
| --- | --- | --- |
| **Credit Rule** | | |
| **action** | **score** | **description** |
| comments | **1** |  |
| make suggestion | **1** |  |
| suggestion got approved | **2** |  |
| edit annotation | **2** |  |
| add annotation | **2** |  |
| edit/add got approved | **2** |  |

### Use Case: Edit Specialty

Each user related to one or more specialties, and each specialty correspond to one specie, and each specie correspond to one or many genes. So each user could only edit some of the genes, as well as adding annotations to them.

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.

## Annotation management subsystem

Keeping track of all annotation changes

The system will keep all the annotations saved no matter they are valid or invalid. So if any person edit one existed annotation record, we will just make this old record invalid anymore and a new annotation will be created waiting for approved.

### Use Case: browse/edit/add 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 want to do. We want users to be able to easily search and manage annotations.

|  |  |  |
| --- | --- | --- |
| **INPUT: edit annotation** | | |
| **NAME** | | **DESCRIPTION** |
| object ID | | the gene being associated |
| ontology type | get from Amigo | e.g. trait ontology |
| term ID | e.g. TO:000001 |
| term name | leaf length |
| evidence code | | select from drop list e.g. IMP |
| score | | select from drop list short/long |
| additional expansion (1-n) | contact relation | select from drop list  e.g. assayed at growth stage/ assayed in mature part |
| term ID | e.g. PO:00000001 |
| term name | e.g. flower stage/leaf |
| evidence | | publication ID |

### Use Case: Save annotation draft

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

### Use Case: Save note

### Use Case: flag annotation

Users can flag an 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

|  |  |
| --- | --- |
| **INPUT: edit annotation** | |
| **NAME** | **DESCRIPTION** |
| object ID | the gene being associated |

### Use Case: Comment on annotation

|  |  |
| --- | --- |
| **INPUT: edit annotation** | |
| **NAME** | **DESCRIPTION** |
| object ID | the gene being associated |

## Object management subsystem

The object here means the object could be associated with the annotations, it could be Gene, Germplasm or QTL.

|  |  |  |
| --- | --- | --- |
| **INPUT: edit gene** | | |
| **NAME** | | **DESCRIPTION** |
| object type | | gene/germplasm/QTL (drop list) |
| object ID | | gene\_ID = gene accession number |
| object name | | gene name  (select from symbol and synonyme) |
| object symbol | | gene symbol |
| object synonyme | | gene synonyme |
| object description | | gene description (free text) |
| object taxon | genus |  |
| species |  |
| genome location | chromosome number |  |
| start | # unit |
| stop | # unit |
| associated publication | |  |
| DB cross reference | source name |  |
| source ID |  |
| URL |  |
|  |  |  |

## Publication management subsystem

|  |  |  |
| --- | --- | --- |
| **INPUT** | | |
| **NAME** | | **DESCRIPTION** |
| publication type | | Journal  Conference  Poster |
|
|
| publication accession ID | | Publication Source ID / Publication DOI |
| publication source name | |  |
| publication source ID | |  |
| publication title | |  |
| publication author | first name |  |
| middle name |  |
| last name |  |
| journal name | |  |
| journal volume | |  |
| journal issue | |  |
| journal paragraph | page start |  |
| page stop |  |

# API design

In this section, all the APIs related to Gene Curation Tool system will be descripted. Since all these API development will be the last task for our project, all the information here may be modified in the future.

## Object import API

The Gene Curation Tool is a public free community which worldwide biologists could utilize it to annotate their own data. Only if they provide us an object data in required format, we could load the data to our database. Then the users could make annotations to these newly added objects (gene, germplasm and QTL).

So, we may need to standardize an object format which require users to follow. And also an API to load the object data.

## Annotation import API

We will provide an API to import annotations from other database. The GAF format is a worldwide accepted Annotation file format, in our system, we will also use this format to load the data.

## Annotation export API

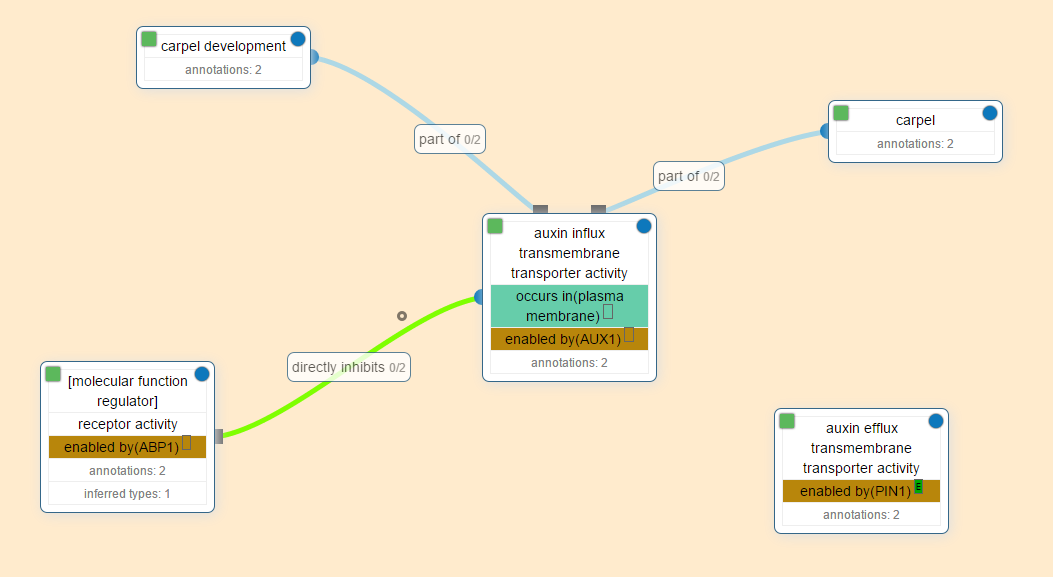
As same as import, we will also provide GAF format file to save the annotation data saved in this system. However, since this system will not only focus on the GO, so the annotation data will also include some different and more enriched information. So we may need to provide another Format which would be accepted by Amigo2 to synchronize the data.

## Utilize API to get Ontology information

Since gene curation tool pay more attention to the association and gene data, and the annotation data will be synchronized with the Amigo2 system at last. So we need to avoid the duplicate data as much as possible. So we will only use ontology ID when saving the association, and then get all the information of the ontology from the Amigo2 system with using the API it provided.

# User Interface Design

TBA…



# Database Design

In this section, we will illustrate the database design of the GCT project.

## ER diagram



Figure 3: 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 many annotations or no annotations, so we call that optional to many.

## Tables design

In this section, we will itemize all tables in the database and the characteristic of rows of the table.

### Table: Users

|  |  |  |  |
| --- | --- | --- | --- |
| **ATTRIBUTE NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| User\_ID | PK | Integers | no |
| 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 or 2 (Defines the level of the user) |  |  |
| Credit | the credit for the contribution of the system |  |  |

### Table: User\_banned

|  |  |  |  |
| --- | --- | --- | --- |
| **ATTRIBUTE NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Index | PK |  |  |
| Banned\_ID | User\_ID |  |  |
| Admin\_ID | Manager who banned the user |  |  |
| Start\_time | Start ban time |  |  |
| End\_time | End ban time |  |  |
| Reason | Reason for banning |  |  |

### Table: Specialty

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Specialty\_ID | Specialty ID PK |  |  |
| Specialty\_name | Specialty Name |  |  |
| Species\_ID | Species ID |  |  |

### Table: User\_Specialty

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Index | PK |  |  |
| User\_id | User ID |  |  |
| Specialty\_id | Specialty ID |  |  |

### Table: Object

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Object\_ID | Gene\_ACC /Unique Gene ID/ PK |  |  |
| Object type | gene/geneplasm/QTL |  |  |
| Object\_name | Gene Name |  |  |
| Object\_symbol | Gene symbol |  |  |
| Object\_description | Gene Description |  |  |
| Chromosome\_number | Genome\_location |  |  |
| Start | Chromosome Start |  |  |
| end | Chromosome End |  |  |

### Table: Synonyms

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Index | PK |  |  |
| Object\_ID | Unique Gene ID |  |  |
| Synonym\_name | Synonym Name |  |  |

### Table: Species

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Species\_id | Species ID PK |  |  |
| Species\_name | Species Name |  |  |
| Genus\_name | Genus\_name |  |  |

### Table: Gene\_Species

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Index | PK |  |  |
| Object\_ID | Gene ID |  |  |
| Species\_id | Species ID |  |  |

### Table: Annotation

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Annotation\_ID | Annotation ID PK |  |  |
| Ontology\_ACC | Ontology ID |  |  |
| Object\_ID | Gene ID |  |  |
| Created\_by | Who added the annotation (user\_ID) |  |  |
| Created\_date | Date submitted |  |  |
| Draft\_flag | 0: in processing 1: submitted |  |  |

### Table: Annotation\_Comment

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Index | PK |  |  |
| Annotation\_ID | Annotation ID PK |  |  |
| Created\_by | Who added the annotation (user\_ID) |  |  |
| Created\_date | Date submitted |  |  |
| Content |  |  |  |

### Table: Annotation\_Note

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Index | PK |  |  |
| Annotation\_ID | Annotation ID |  |  |
| Created\_by | Who added the annotation (user\_ID) |  |  |
| Created\_date | Date submitted |  |  |
| Content |  |  |  |

### Table: Annotation\_Validation

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Index | PK |  |  |
| Annotation\_ID | Annotation ID |  |  |
| Valided\_by | User ID |  |  |
| isValid | -1 invalid or 1 valid |  |  |
| Valided\_date | Date |  |  |
| Reason | Reason why flagged |  |  |

### Table: Annotation\_Approvement

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Index | PK |  |  |
| Annotation\_ID | Annotation ID |  |  |
| Approved\_by | User ID |  |  |
| isApprove | -1 disapprove or 1 approve |  |  |
| Approved\_date | Date |  |  |
| reason | Reason why |  |  |

### Table: Approved\_Annotations

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Index | PK |  |  |
| Annotation\_ID | Annotation ID |  |  |
| Approvement\_ID |  |  |  |

### Table: Evidence

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Evidence\_ID | Evidence ID PK |  |  |
| Evidence\_Code | Evidence Code |  |  |

### Table: Annotation\_Evidence

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Index | PK |  |  |
| Annotation\_ID | Annotation ID |  |  |
| Evidenc\_ID | Evidence ID |  |  |

### Table: Publications

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Publication\_ID | Publication ID |  |  |
| Publication\_type |  |  |  |
| Publication\_source\_name |  |  |  |
| Publication\_source\_ID |  |  |  |
| Publication\_title | The name of publication |  |  |
| Publication\_DOI | Source |  |  |
| Abstract | Abstract Text |  |  |

### Table: Author

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Author\_ID | Author ID |  |  |
| Author\_first\_name | Author’s First Name |  |  |
| Author\_last\_name | Author’s Last Name |  |  |

### Table: Author\_Publication

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Author\_ID | Author ID |  |  |
| Publication\_ID | Publication ID |  |  |

### Table: Xref

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| Source\_name | Source Name |  |  |
| Source\_ ID | Unique ID for Xref\_Object in the source |  |  |
| URL | Reference to Ontology ACC |  |  |

### Table: Xreference\_relation

|  |  |  |  |
| --- | --- | --- | --- |
| **ROW NAME** | **DESCRIPTION** | **DATA TYPE** | **NULLABLE** |
| XRef\_ID | XRef ID |  |  |
| Object\_ID | Object\_ID |  |  |

# References

1. Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis et al. "Gene Ontology: tool for the unification of biology." Nature genetics 25, no. 1 (2000): 25-29.<http://www.nature.com/ng/journal/v25/n1/full/ng0500_25.html>
2. Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." Nucleic acids research 37, no. 1 (2009): 1-13.
3. Du, Zhou, Xin Zhou, Yi Ling, Zhenhai Zhang, and Zhen Su. "agriGO: a GO analysis toolkit for the agricultural community." Nucleic acids research(2010): gkq310.
4. The Planteome project proposal
5. iPlant<http://www.iplantcollaborative.org/about-iplant>
6. Planteome Project Proposal<http://wiki.planteome.org/images/1/1d/Planteome_grant_NSF_-1340112.pdf>
7. G8 Planteome Presentation Slides [G8\_open\_data\_2013\_DC.pdf](https://www.dropbox.com/s/nok7gilog2t5id7/Jaiswal_cROP_G8_open_data_2013_DC.pdf)
8. Genome definition<https://en.wikipedia.org/wiki/Genome>
9. Phenotype definition<https://en.wikipedia.org/wiki/Phenotype>
10. Gene Ontology [http://biochem218.stanford.edu/Projects%202010/Blair%202010.pdf](%20http://biochem218.stanford.edu/Projects%202010/Blair%202010.pdf)
11. Protein definition<https://en.wikipedia.org/wiki/Protein>
12. Some concepts on ontologies, semantic nets, and taxonomy<http://ascelibrary.org/doi/pdf/10.1061/(ASCE)0887-3801(2005)19:4(394)>
13. History of Ontology Info<http://www.cs.vassar.edu/~weltyc/papers/fois-intro.pdf>
14. 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>
15. Gene Ontology Annotations<http://nar.oxfordjournals.org/content/37/suppl_1/D555.full.pdf+html>
16. Ontologies in Relational Databases<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.2534&rep=rep1&type=pdf>
17. Jaiswal, Pankaj, et al. "Gramene: development and integration of trait and gene ontologies for rice."*Comparative and functional genomics* 3.2 (2002): 132-136.<http://onlinelibrary.wiley.com/doi/10.1002/cfg.156/epdf>
18. Lens, Frederic, et al. "An extension of the Plant Ontology project supporting wood anatomy and development research." *IAWA Journal* 33.2 (2012): 113-117.<http://www.researchgate.net/publication/232271560_An_extension_of_the_Plant_Ontology_project_supporting_wood_anatomy_and_development_research>

# Appendix A: Definitions and Abbreviations

|  |  |
| --- | --- |
| GCT | Gene Curation Tool |
| cROP | “Common Reference Ontology for Plants”, a set of ontologies concerning plants. This was the name of the project before it was changed to Planteome by Dr. Pankaj. |
| Ontology | 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. Ultimately, an ontology can be a vital tool enabling researchers to turn data into knowledge. |
| Annotation | 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. |
| 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). [1] |
| AmiGO | A web tool for accessing the Gene Ontology project’s data (including browsing genes and their corresponding annotations). |
| Genomes | The complete genetic material of an organism [11]. |
| MapReduce | A programming model that where you split up data across processes, so that they can be ran independently in parallel [6]. |
| NoSQL | “Not Only SQL”, steer away from your traditional relational database model for better performance on flat data [6]. |
| Phenotypes | Observational characteristics of an organism [11]. |
| Protein | Large [biomolecules](https://en.wikipedia.org/wiki/Biomolecule), or [macromolecules](https://en.wikipedia.org/wiki/Macromolecule), consisting of one or more long chains of [amino acid](https://en.wikipedia.org/wiki/Amino_acid) [residues](https://en.wikipedia.org/wiki/Residue_(biochemistry))” [8]. |