Gene Curation Tool (GCT)

Design Document

Data and Visualization Group

Botong Qu

Jaden Diefenbaugh

Eugene Zhang

August 2015

Contents

1 Introduction

The Planteome Project (http://www.planteome.org), a three year plan, is an international collaboration to support the development of "Common Reference Ontologies and Applications for Plant Biology" (cROP). By giving researchers an easier way to collaborate with one another and by giving them an easy way to search through and manage annotations, the project will speed up the overall ontological annotation research process.

1.1 Planteome Project

The importance of Planteome

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]. In this case, scientists keep trying to produce crops which are more tolerant or resistant to the drought, temperature, diseases and nutrient deficiencies. Traditional plant breeding methods alone are not sufficient and efficient to overcome these challenges, but new methods such as high-throughput sequencing and automated scoring of phenotypes can provide assistance in the form of significant new insights in plant biology [4].

There are two main types of genomics-assisted breeding [20]: (1) MAS and (2) GS. MAS, uses molecular markers that map within specific genes or QTLs known to be associated with target traits or phenotypes to select individuals that carry favorable alleles for traits of interest (and/or to discard those that do not). GS, on the other hand, uses all available marker data for a population as predictors of breeding value. The advantage of genomics-assisted breeding is that genotypic data obtained from a seed or seedling can be used to predict the phenotypic performance of mature individuals without the need for extensive phenotypic evaluation over years and environments. The use of genomics-assisted breeding, in both MAS and GS, allows for more selection cycles and greater genetic gain per unit of time [19].

There are lots of sequenced genomes of Viridiplantae species being available now, such as Phytozome database, Grammene database. The growing of the high-throughput phenotype and diversity date is rapid. Analyses of these vast data from genetic and genomic studies have the potential to improve our understanding of species evolution (how different organisms are related and how they evolved), development and the molecular basis of economically relevant traits.

The problems we try to solve

In order to utilize all genetic data around the world, researchers must be able to connect the spatial and temporal expression patterns of genes and gene products to their molecular functions, and elucidate their roles in biological processes and potential gene-

gene interactions. Effective interspecies comparison demand a common, relational vocabulary an ontology, to permit computer-aided reasoning on biological entities genomic scale.

Goals of Planteome

Planteome will 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. Beside, we will also 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. Our hope is to speed up the plant research process by taking advantage of above methods that 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.

cROP

The Planteome 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.

1.2 Gene Curation Tool (GCT)

Definition of Gene Curation

Data curation is a term used to indicate management activities required to maintain research data long-term such that it is available for reuse and preservation. In science, data curation may indicate the process of extraction of important information from scientific texts, such as research articles by experts, to be converted into an electronic format, such as an entry of a biological database. In broad terms, curation means a range of activities and processes done to create, manage, maintain, and validate a component.

A biocurator is a professional scientist who curates, collects, annotates, and validates information that is disseminated by biological and model organism databases. The role of a biocurator encompasses quality control of primary biological research data intended for publication, extracting and organizing data from original scientific literature, and describing the data with standard annotation protocols and vocabularies that enable powerful queries and biological database inter-operability. Biocurators communicate with researchers to ensure the accuracy of curated information and to foster data exchanges with research laboratories [21].

In genome annotation, for example, biocurators commonly employ—and take part in the creation and development of—shared biomedical ontologies: structured, controlled vocabularies that encompass many biological knowledge domains, such as the Open Biomedical Ontologies found in the OBO Foundry. These domains include genomics and proteomics, anatomy, animal and plant development, biochemistry, metabolic pathways, taxonomic classification, and mutant phenotypes [21].

What is the purpose of developing 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.

Traditionally, biological knowledge has been aggregated through expert curation, conducted manually by dedicated experts. However, with the burgeoning volume of biological data and increasingly diverse densely informative published literatures, expert curation becomes more and more laborious and time-consuming, increasingly lagging behind knowledge creation. Community Curation harnesses community intelligence in knowledge curation, bears great promise in dealing with the flood of biological knowledge.

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.

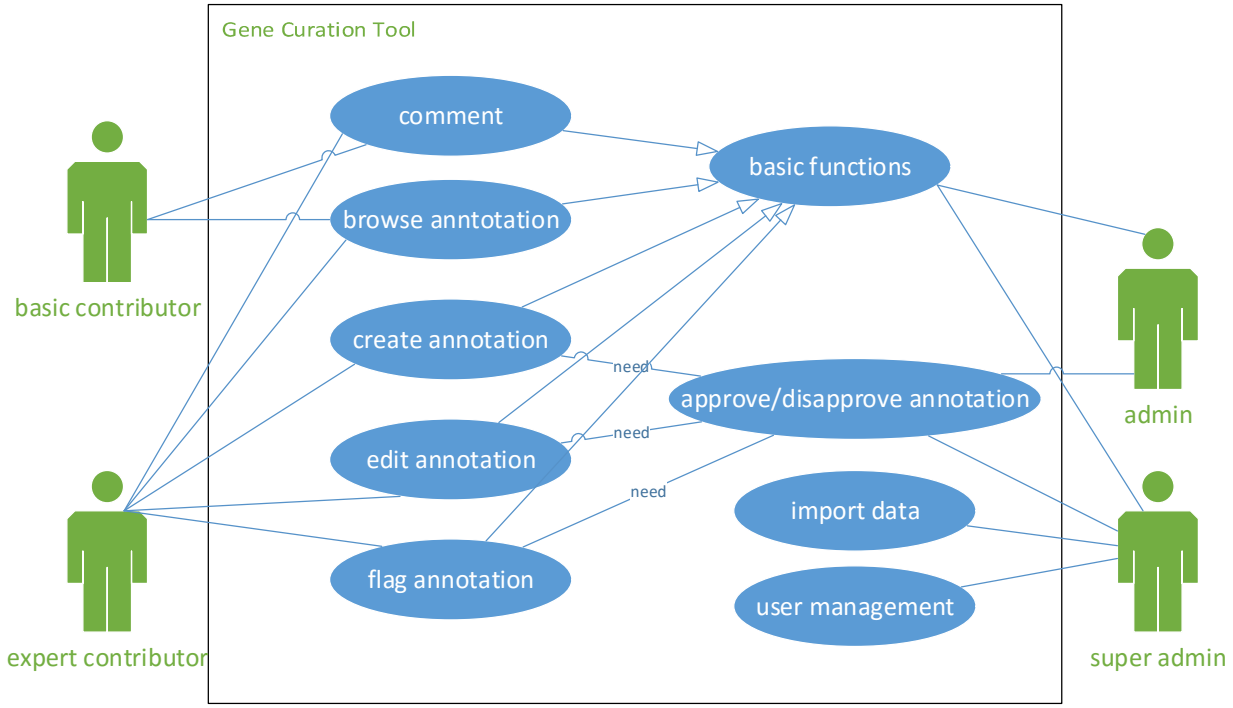


Figure 1: This diagram shows the basic process of creating, editing, and flagging (e.g. All three of these actors can create, edit, and flag annotations, but only the admin and super admin can approve/disapprove an annotation). One can also see that each time an annotation is edited, flagged (a certain 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 an admin or an super admin with a specialty in that area.

The meaning of GCT.

This will be the first such resource hosting a complex set of ontologies and the associated annotations in any bioinformatics project. Therefore, this pilot project will be instrumental in driving the development of new scalable data stores and user interface for all of biology, not just plants.

1.3 Features of GCT

Based on above statement, we conclude following features of GCT:

Unlike Amigo which only build based on Gene Ontology (Amigo2 is extending to other ontology data), GCT will provide all ontology data access to the contributors of the community. This feature is also agree with the development of the cROP. Scientists come from all around the world need a central platform the share, query and browse the Effective interspecies genomic annotations. This would facilitate them an effective and convenient way to make use of all the genomic sequencing result.

GCT will also act as a backup of Amigo2 database, which means that GCT database could be easily modified, annotated, curated, meanwhile keep the Amigo2 database safe and accuracy. Opening the annotation database to the whole community is a risky

job, the high quantity of all biocurators from the world and their contributions to the database would easily lead to reduction of the accuracy, and may also lead to conflicts between the conclusions. So a certain work on management of these data would be necessary. In GCT, the admin and super admin who will be able to approve or disapprove the annotation edition would guarantee this feature.

GCT would also become the most comprehensive annotation database, not only keep the approved data, but also contain all history of the modification. The GCT database would save the data in process, unapproved, out of date. Keeping all these history would be helpful for the users to understand the evolvement of the annotation data. And may also be easy for the admins to backtrack some curation.

GCT would perform as a communication platform for the whole community. Users are designed to be able to comment on the annotation data and make suggestion about information being shown. All these action would convey abundant information to other users and administers of the system. Besides, the credit mechanism would also be able to encourage the contribution among the community members.

1.4 Basic knowledge about Genes

We will introduce some basic genetic knowledge to the readers for a better understanding of the whole project. 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 or 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).

1.5 Ontology

What is ontology?

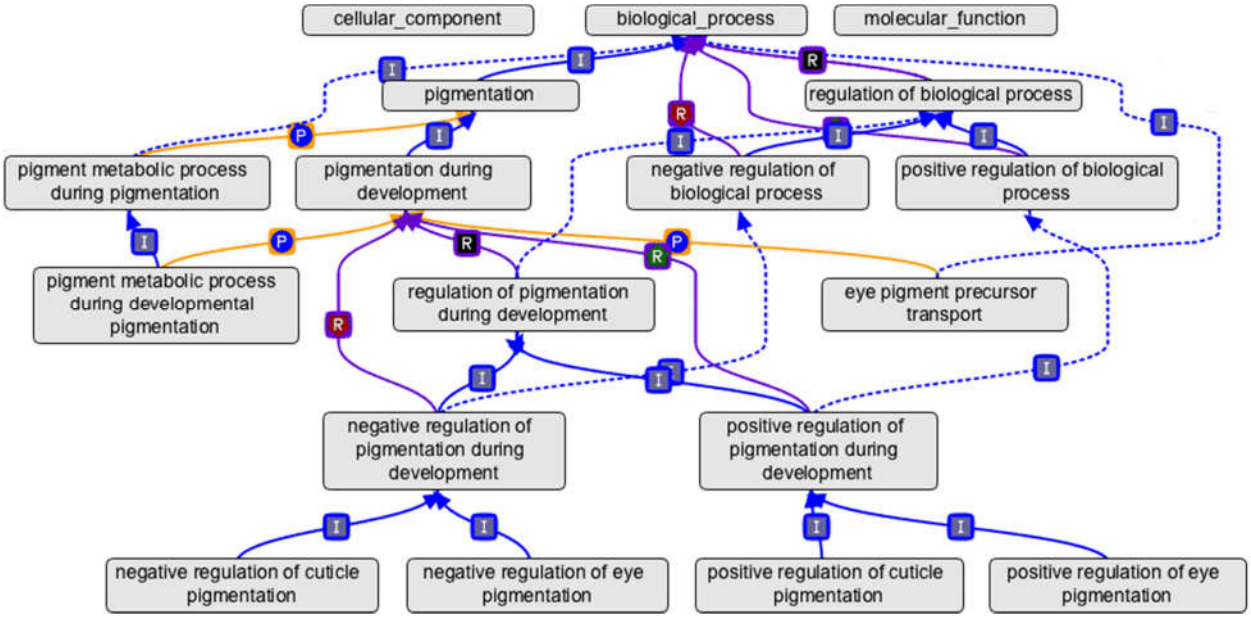
Ontology is the philosophical study of the nature of being, becoming, existence, or reality, as well as the basic categories of being and their relations, can be dated as far back as 1613, its practice however can be dated as far back as to Aristotle. Philosophical ontology has sought the definitive and exhaustive classification of entities in all spheres of being.

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 knowledge in specific domain 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.

Ontology Chart/Graph

An ontology chart is a type of chart used in semiotics and software engineering to illustrate an ontology. For example. The structure of GO can be described in terms of a graph, where each GO term is a node, and the relationships between the terms are edges between the nodes. GO is loosely hierarchical, with 'child' terms being more specialized than their 'parent' terms, but unlike a strict hierarchy, a term may have more than one parent term (note that the parent/child model does not hold true for all types of relation, see the relations documentation). For example, the biological process term hexose biosynthetic process has two parents, hexose metabolic process and monosaccharide biosynthetic process. This is because biosynthetic process is a subtype of metabolic process and a hexose is a subtype of monosaccharide.

The following diagram is a screenshot from the ontology editing software OBO-Edit, showing a small set of terms from the ontology.



In the diagram, relations between the terms are represented by the colored arrows; the letter in the box midway along each arrow is the relationship type. Note that the terms get more specialized going down the graph, with the most general terms—the root nodes, cellular component, biological process and molecular function—at the top of the graph. Terms may have more than one parent, and they may be connected to parent terms via different relations. The GO relations documentation describes these relations in greater detail.

Ontology in biology

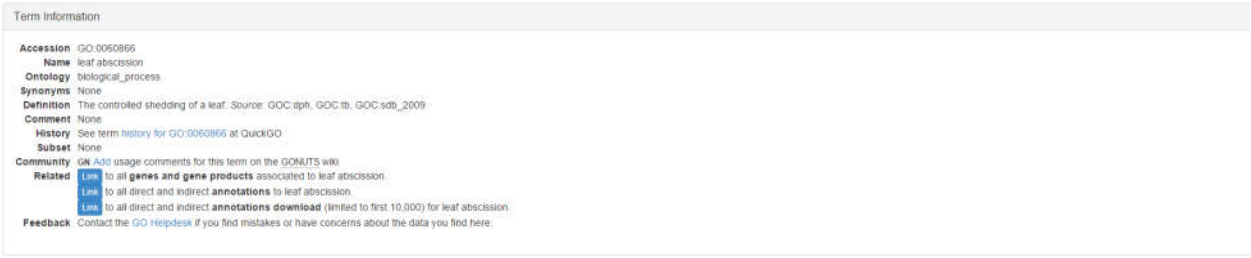
In biology, a key element of a semantic data integration framework is the ontology, a formal representation of a knowledge domain in which concepts, terms and classes are interrelated in a graph or network, with edges indicating relationships between these concepts. Data are attached to terms through 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.

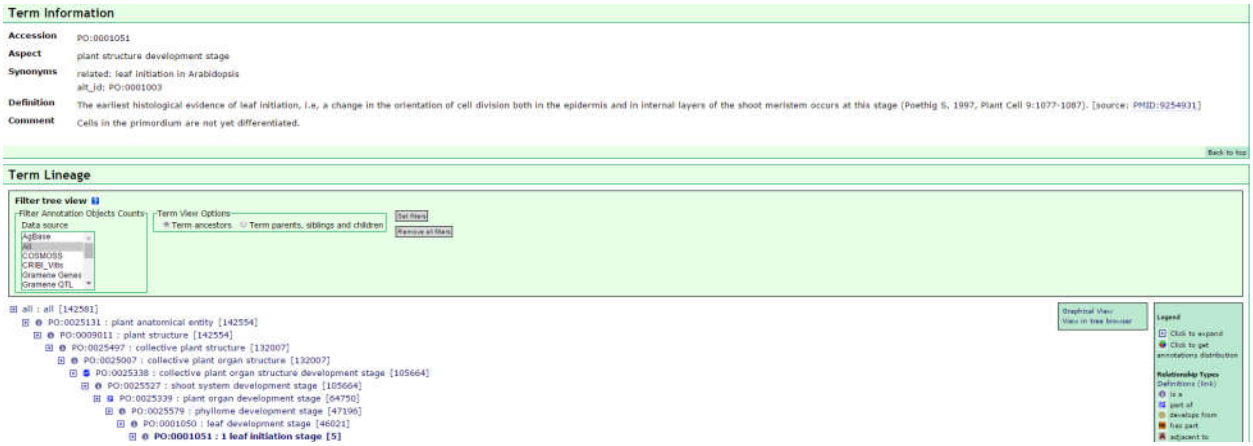
For example, one could use GO to conclude that a specific protein found in eukaryotic cells relates to “core biological process” and is common among all eukaryotic cells [1]. Furthermore, one can see that understanding how genes among various species and how they relate can be vital when trying to understand 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.

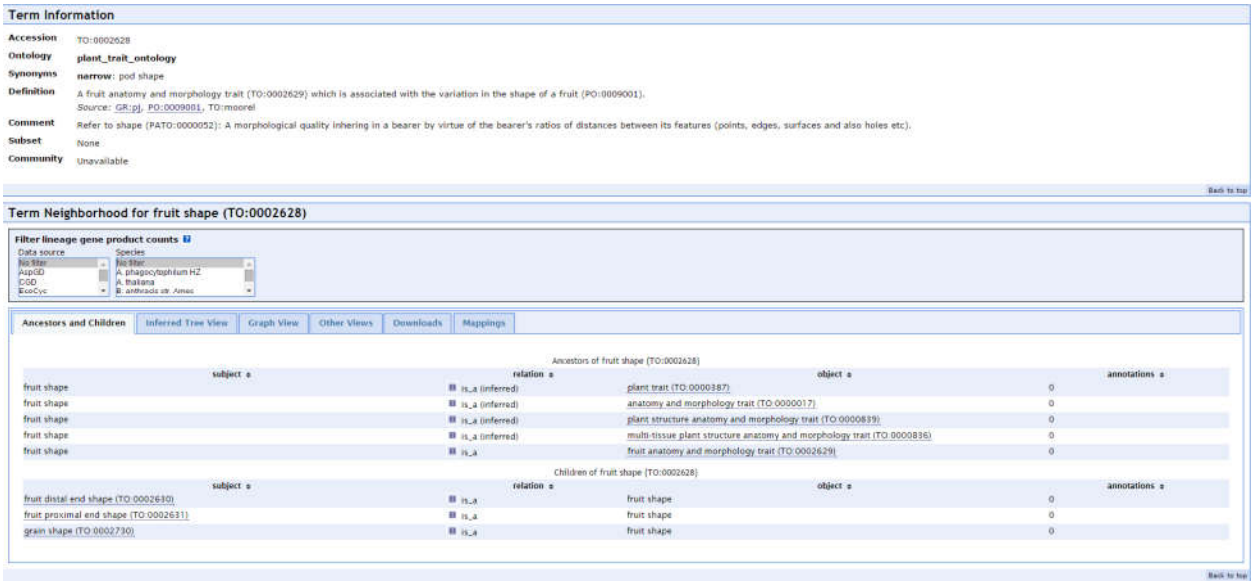
For example, PO: 0001051 is the *leaf initiation stage*, its aspect is plant structure development stage. Its definition is: the earliest histological evidence of leaf initiation, i.e, a change in the orientation of cell division both in the epidermis and in internal layers of the shoot meristem occurs at this stage (Poethig S, 1997, Plant Cell 9:1077-1087). Just like other ontologies it uses relationships (e.g. is_a and part_of are the most common) to interpret how things relate [19]. In this example, we know that the *leaf initiation stage* is a leaf development stage (PO: 0001050).



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.

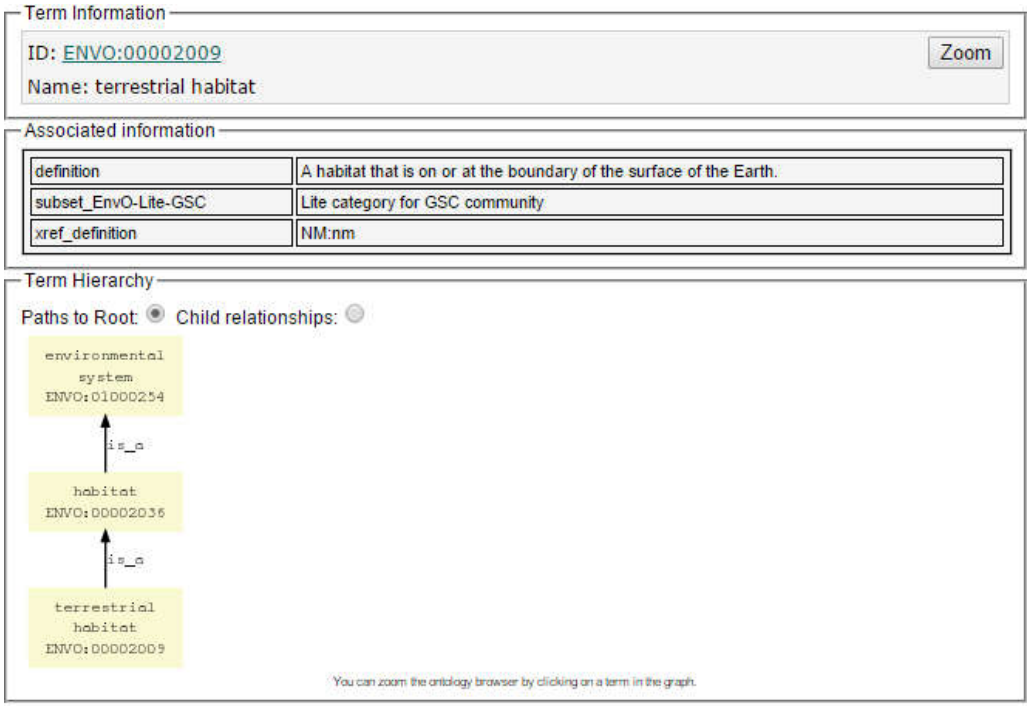
The TO is a 'composite' or 'pre-composed' ontology, meaning the terms are assembled in advance from simpler 'building block' concepts. For example, the trait *fruit shape* (TO: 0002628) is an observable characteristic composed using the Plant Ontology term fruit (PO: 0009001) and the PATO attribute (quality) 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. From the definition, we could find that this ontology is_a fruit anatomy and morphology trait (TO: 0002629) which is associated with the variation in the shape of a fruit (PO: 0009001).



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.

For example, greenhouse soil (ENVO: 00005780) is_a soil (ENVO: 00001998), soil is part_of terrestrial habitat (ENVO: 00002009) which is a habitat (ENVO: 00002036).



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.

1.6 Annotation

What is Annotation?

Annotation is the process of assigning ontology terms to gene or genetic resources such as germplasm or QTL. Information about a gene that is attached to these vocabularies (concepts) in ontologies and used to describe their relationships. An annotation often contain an evidence code and literature associated with it to back up this newly found information about a gene according to Dr. Pankaj.

GO Annotation examples

To clarify this definition, we still use the GO as an example. The annotation data in the GO database is contributed by members of the GO Consortium, and the Consortium is continuously encouraging new groups to start contributing their annotations.

Negative regulation of leaf senescence (GO:1900056) is an ontology of biological process in GO database. GLK2 (Gene/product name: AT5G44190) is a gene which had been associated to Negative regulation of leaf senescence by UniProtKB at May 28th 2015. The evidence is IMP and refer to TAIR: Publication:501754151 and PMID:23459204. One can search all these ontological annotation from databases such as Amigo.

Gene/product	Gene/product name	Qualifier	Direct annotation	Annotation extension	Assigned by	Taxon	Evidence	Evidence with	PANTHER family	Isoform	Reference	Date
GLK2	AT5G44190		negative regulation of leaf senescence		UniProtKB	Arabidopsis thaliana	IMP			TAIR:locus:2167639	TAIR:Publication:501754151 PMID:23459204	20150528

Where the Annotation data come from?

Annotations can be specified by a curator selecting a term from lower or higher up the hierarchy. Annotations are made to ontologies 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 ontology also changes as terms are added, or removed, annotations are therefore updated periodically.

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.

The use of Annotation in Bioinformatics

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data.

Since 1977, the DNA sequences of thousands of organisms have been decoded and stored in databases. This sequence information is analyzed to determine genes that encode proteins, RNA genes, regulatory sequences, structural motifs, and repetitive sequences. A comparison of genes within a species or between different species can show similarities between protein functions, or relations between species (the use of molecular systematics to construct phylogenetic trees). These research of the finding relationship would be useful in breading featured plant and curing the disease.

Annotation is one aspect of bioinformatics in sequence analysis. In the context of genomics, annotation is the process of marking the genes and other biological features

in a DNA sequence. This involves computational gene finding to search for protein-coding genes, RNA genes, and other functional sequences within a genome. Not all of the nucleotides within a genome are part of genes. Recent software patented are now able to identify tissue-specific alternative splicing events, which will allow scientist to identify multiple gene products that come from the same transcripts. In that way, Bioinformatics helps to bridge the gap between genome and proteome projects — for example, in the use of DNA sequences for protein identification.

1.7 About this document

Why does GCT need a design document?

The Gene Curation Tool (GCT) development of Planteome is a long term project, which currently has been planned to be finished in next three years, and would need a great of cooperation between lots of people come with different background knowledge and specialties.

When developing a large-scale project like Planteome, any little misunderstanding and indistinct definition appear among the group would possibly lead to a huge gap, which could result in that the developers create something totally opposite from the original purpose. If this happened, the waste of time and manpower would seriously delay the progress of the whole project. Besides, the long term project always facing the member changing problem, tutoring new comers to help them quickly grasp the big picture of the project always takes a lot of time.

In this case, a detailed design document which specifies all related information and knowledge of the project, would be helpful for all the participants, no matter they specialize in biology or computer science, to comprehend the whole project. And this authoritative handbook and reference would help all related member be synchronized while developing, for all design and understanding would be based on and come from one common resource. For all above reasons, a carefully and throughout reading of this document becomes not only recommended, but also necessary.

The structure of this document and guide of reading.

This document is intended for people developing and/or supervising the Gene Curation Tool (GCT) project, and could be divided into two parts, the introduction 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.

2 Existing 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.

2.1 Amigo

Amigo is closely related to our project. In this section, we will review Amigo and explain how we plan to interact with it in our project.

2.1.1 Introduction

Amigo is a tool used for the viewing of an ontology graph (relations between terms) and has capabilities for suggesting cross-references and researching annotations. It is a web-based application that allows users to query, browse and visualize ontologies and gene product annotation data. In addition, it also has a BLAST tool, tools allowing analysis of larger data sets, and an interface to query the GO database directly. AmiGO can be used online at the GO website to access the data provided by the GO Consortium (GOC), or can be downloaded and installed for local use on any database employing the GO database schema. It is free open source software and is available as part of the go-dev software distribution. AmiGO2 is a project to create the next generation of AmiGO---the current official web-based set of tools for searching and browsing the Gene Ontology database.

Gene Ontology Consortium (GOC) is a collaborative effort to address the need for consistent descriptions of gene products across databases, and has continued following this goal statement by developing tools to those looking to computational analysis gene data. It is the set of biological databases and research groups actively involved in the gene ontology project. This includes a number of model organism databases and multi-species protein databases, software development groups, and a dedicated editorial office.

The term ***gene ontology*** is suitably nebulous to describe the domain of both the GOC and AmiGO, but does little to explain itself unless one has a biology and philosophy background. To explain, the original problem will be given, and then an explanation of how a gene ontology solves it.

In five words, “fragmentation and lack of standards” may describe the problem the best. Researching the building blocks of life, genes, has been an increasingly complex affair. Many groups, from nations to universities, have been researching genes and compiling their own data and analyses. While progress was being made, sharing and collaborating became an issue - different formats and conventions meant that too much time was spent adjusting existing workflows, tools, and databases to be more open and well-designed so that others could use them as well. People wanted to be able to work on data from different groups and organisms seamlessly; genetically, much of the life on earth is similar - therefore the depiction of all this data should, logically, be similar as well.

Realizing that genetic data was similar enough to be described in the same ways, the GOC developed an ontology of this data. An ontology is a vocabulary of a domain. It takes all the concepts for a topic (genes, in this case) and labels all these concepts, along with defining relationships between these labels. Before this gene ontology was developed, everyone was creating their own labels, descriptions, and relationships of their gene data; each group developed their own gene ontology. Now, the GOC has developed and standardized an ontology for genes, and their goal is to format as much data as they can using their ontology. This means that anyone who is familiar with one of the GOC’s datasets can interact with all the datasets with little friction, because they are all described the same way. AmiGO is a system that interacts with this data they have converted (stored in the Gene Ontology database), and allows users to easily search this trove of gene data for their own research.

2.1.2 Features

AmiGO 2 is a front-end database explorer and analyzer, primarily concerning itself with the gene ontology terms, gene data, and *annotations*. Annotations, as defined by AmiGO, are “associations between GO terms and genes or gene products.” They bind all the gene data in the GO database together, allowing for cause-and-effect studies, history, and other research that involves relationships.

The AmiGO 2’s most prominent features include the following:

Quick Search

Simple, search engine-inspired, single-line text input for fuzzy-matching from the database’s string fields

Grebe Search Wizard

Similarly simple, fill-in-the-blank searcher with common prompts to help choose where to search in the database

Advanced (Lucene) Search

Split into three searchers for each major type of data (“Annotations”, “Ontology”, “Gene and gene products”), this supports powerful live searches within one of the three topics, and includes boolean operator support and (limited) regex support

GOOSE

The “GO Online SQL Environment” directly searches the GO (LEAD) database and its mirrors via user-submitted SQL

Term Enrichment Service

Performs “enrichment analysis” on the GO gene data. Enrichment analysis is analysis concerning how “rich” a specific term’s data is in the GO database - whether there is “enough” data about a term and its conditions. One may think of this as limited meta-analysis.

Statistics

A static page showing various database statistics, generated on Amazon’s AWS3 cloud.

Custom Visualization

Generate a static visualization of the data of one’s choice; the basic version has users entering GO IDs (the unique ID of a piece of data) and the AWS3 cloud generates an image, while a much more advanced version is able to used structured JSON (a simple object notation specification comparable to XML) for images.

Others

The AmiGO 2 website holds a possibly exhaustive list of separate software that works on top of AmiGO.

2.1.3 GO Annotation File (GAF) Format

The Gene Ontology Consortium stores annotation data, the representation of gene product attributes using GO terms, in tab-delimited plain text files. Each line in the file represents a single association between a gene product and a GO term with a certain evidence code and the reference to support the link. Annotation data is submitted to the GO Consortium in the form of gene association files, or GAFs. Since we will connect with Amigo a lot, it is essential for us to understand the GAF format be used.

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

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.*

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 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 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 wither 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 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.

Examples

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

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

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. 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 is usually 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 can be used in the **with** column for ISS annotations.

The **with** column may **not** be used with the evidence codes IDA, TAS, NAS, or ND.

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

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.

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 for details of practical usage; a wider discussion of the annotation extension column can be found on the GO wiki.

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 includeUniProtKB accession numbers, NCBI NP identifiers or Protein Ontology (PRO) 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.

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)

2.2 Gramene

Gramene is a curated, open-source, integrated data resource for comparative functional genomics in crops and model plant species. Our goal is to facilitate the study of cross-species comparisons using information generated from projects supported by public funds. Gramene currently hosts annotated whole genomes in over two dozen plant species and partial assemblies for almost a dozen wild rice species in the Ensembl browser, genetic and physical maps with genes, ESTs and QTLs locations, genetic diversity data sets, structure-function analysis of proteins, plant pathways databases (BioCyc and Plant Reactome platforms), and descriptions of phenotypic traits and mutations.

Extensive research over the past two decades has shown there is a remarkably consistent conservation of gene order within large segments of linkage groups in agriculturally important grasses such as rice, maize, sorghum, barley, oats, wheat, and rye. Grass genomes are substantially colinear at both large and short scales with each other, opening the possibility of using syntenic relationships to rapidly isolate and characterize homologues in maize, wheat, barley and sorghum.

As an information resource, Gramene's purpose is to provide added value to data sets available within the public sector, which will facilitate researchers' ability to understand the grass genomes and take advantage of genomic sequence known in one species for identifying and understanding corresponding genes, pathways and phenotypes in other grass species. This is achieved by building automated and curated relationships between cereals for both sequence and biology. The automated and curated relationships are queried and displayed using controlled vocabularies and web-based displays. The controlled vocabularies (Ontologies), currently being used include Gene ontology, Plant ontology, Trait ontology, Environment ontology and Gramene Taxonomy ontology. The web-based displays for phenotypes include the Genes and Quantitative Trait Loci (QTL) modules. Sequence based relationships are displayed in the Genomes module using the genome browser adapted from Ensembl, in the Maps module using the comparative map viewer (CMap) from GMOD, and in the Proteins module displays. BLAST is used to search for similar sequences. Literature supporting all the above data is organized in the Literature database.

2.3 AgriGO

The AgriGO is a web-based tool and database for the gene ontology analysis. It supports specicial focus on agricultural species and is user-friendly. The AgriGO is designed to provide deep support to agricultural community in the realm of ontology analysis. Compared to other available GO analysis tools, unique advantages and features of AgriGO are:

1. The AgriGO especially focuses on agricultural species. It supports 45 species and 292 datatypes currently. And AgriGO is designed as an user-friendly web server.

2. New tools including PAGE (Parametric Analysis of Gene set Enrichment), BLAST4ID (Transfer IDs by BLAST) and SEACOMPARE (Cross comparison of SEA) were developed. The arrival of these tools provides users with possibilities for data mining and systematic result exploration and will allow better data analysis and interpretation.

3. The exploratory capability and result visualization are enhanced. Results are provided in different formats: HTML tables, tabulated text files, hierarchical tree graphs, and flash bar graphs.

4. In AgriGO, PAGE and SEACOMPARE can be used to carry out cross-comparisons of results derived from different data sets, which is very important when studying multiple groups of experiments, such as in time-course research.

3 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.

3.1 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.

3.2 Scope

We descript the features in scope of GCT.

a. Design a database will merge the many branches of ontology.
b. Users could easily search and browse annotations associated with a gene, an ontology, etc.
c. Users could easily search and browse information for genes and ontologies, publications, authors, etc.
d. 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.
e. GCT will also works similar to a forum. That is people could easily comment on the existed annotations and communicate with others.
f. 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.

3.3 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

3.4 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.

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 table 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.

3.5 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**	

3.6 Product Functions

GCT should support the following use cases:

Class of use cases	RS_ID	Use cases	description
	1-1	register	

unregistered user's capabilities (basic contributor)	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)	2-1	Login	
	2-2	profile management (include change password)	
	1-2,1-3,1-4	browse annotation/ontology/gene information	
	2-3	export annotation information	
	2-4	browse credit	
	2-5	edit annotation	
	2-6	save annotation draft	
	2-7	save note to user self	
	2-8	flag annotation	
	2-9	add annotation	
	1-5	comment on annotation	

Admin's capabilities	1-1 to 1-5 2-1 to 2-9	all capabilities of registered user	
	3-1	manage user's information (credit, password, profile)	
	3-2	manage users' role	
	3-3	ban user from activities	
	3-4	approve annotation modifications (edit and add)	
	3-5	approve annotation suggestion (flag)	
	3-6	edit publication	
	3-7	edit object	
Super admin's capabilities	1-1 to 1-5 2-1 to 2-9 3-1 to 3-7	All capabilities of admin	
	4-1	Import data (gene data, annotation data, etc.)	

3.7 User management subsystem

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

3.7.1 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	

3.7.2 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: login	
NAME	**DESCRIPTION**
success	
wrong password	
no username	

3.7.3 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 persistent 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

3.7.4 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.

3.7.5 Use Case: Role management

USE CASE: Role management		
Description	Admins and super admins could manage the role of registered users.	
Main Actor	Admins and Super admins	
Trigger	Click the role management button.	
Typical case Scenario	**Action**	**Response**
Alternate Scenario		
Result		
Constraint		

3.8 Annotation management subsystem

The GCT will keep track of the modification history all annotations. If any people edit one existed annotation record, we will create a new record and wait for approvement. If the new record is approved, the old annotation will be flagged as invalid and the new record will be flagged as valid.

3.8.1 Use Case: browse/edit/add annotation

We need to provide an efficient way 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/add 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

3.8.2 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

3.8.3 Use Case: Save note

When user edit the annotation, they could save a note which is only available to him/herself.

INPUT: save note	
NAME	**DESCRIPTION**
Annotation ID	The ID of the annotation to be noted
The author	

date	
Content	

3.8.4 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: flag annotation	
NAME	**DESCRIPTION**
Annotation ID	the annotation being flagged
Date	
Author	
Flag to	Invalid/ valid
reason	

3.8.5 Use Case: Comment on annotation

All users (unregisterd or registered) should be able to comment on annotations, and all these comments are available to all the users. The admin should be able to delete these comments.

INPUT: comment on annotation	
NAME	**DESCRIPTION**
Annotation ID	the annotation being comment

author	
date	
content	

3.9 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	

3.10 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	

4 API design

An API is a protocol intended to be used as an interface by software components to communicate with each other. Publicly available APIs for both internal and external data access will be developed in project. It will open access to the web portal to 3rd party web sites or software of users who want to extract the ontology terms and annotations for their final local re-use.

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.

4.1 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.

4.2 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.

4.3 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

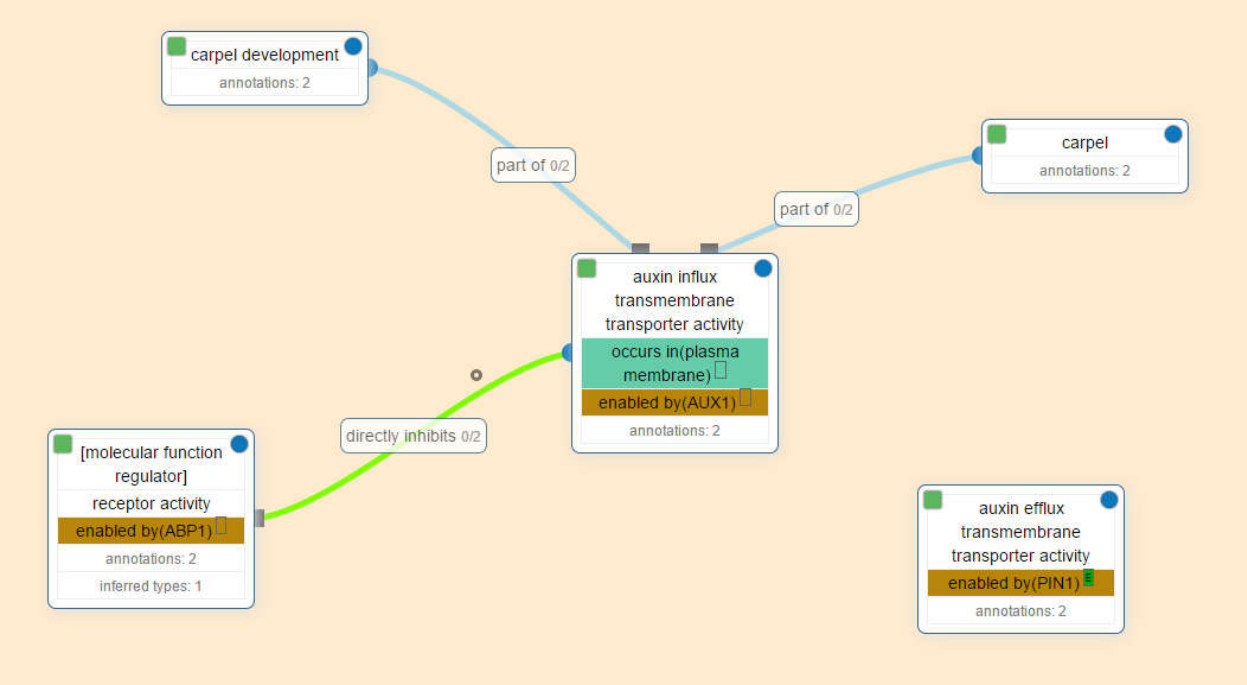
annotation data will also include some different and more enriched information. So we may need to provide another Format which would be accepted by Amigo2 to synchronize the data.

4.4 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.

5 User Interface Design

TBA…



6 Database Design

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

6.1 ER diagram

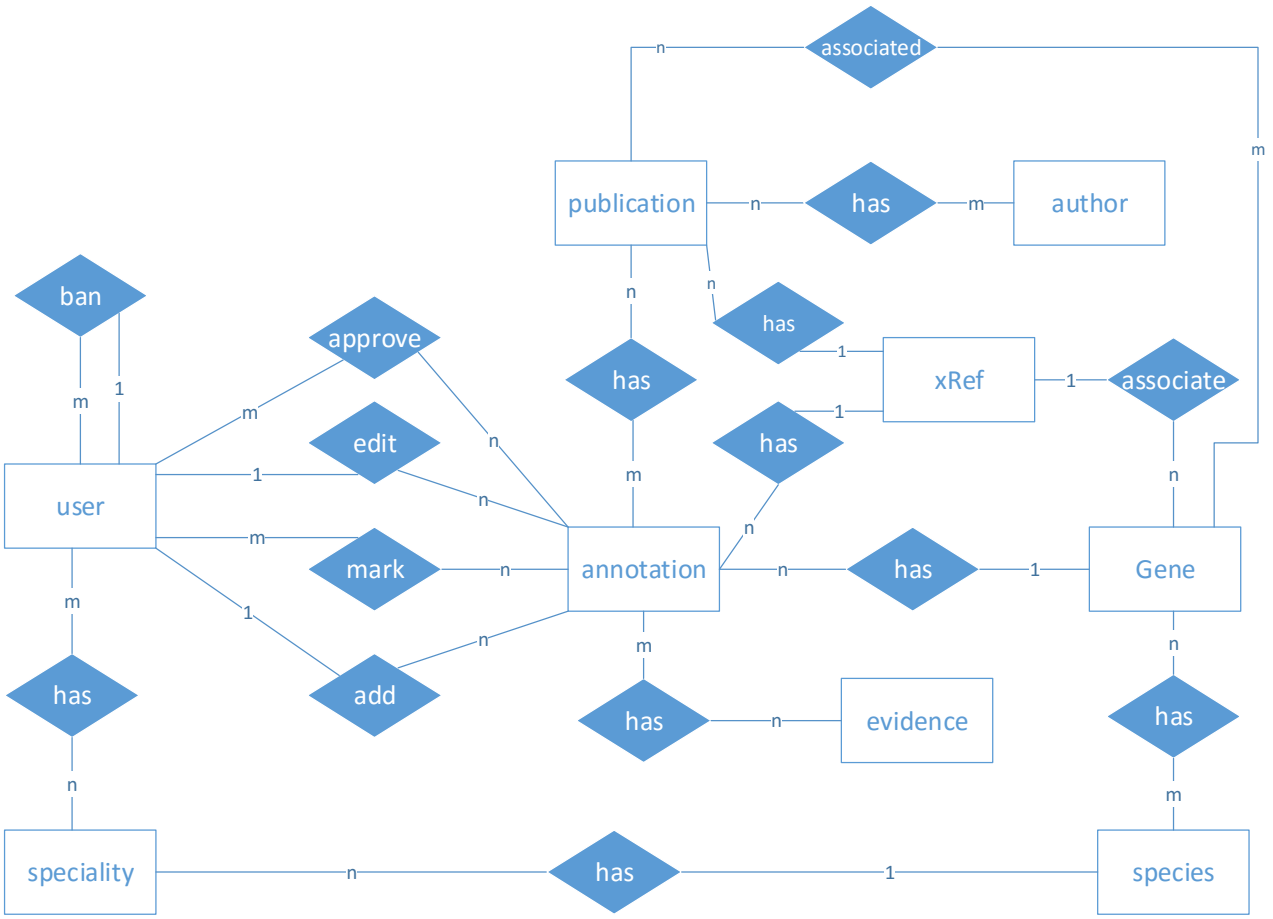
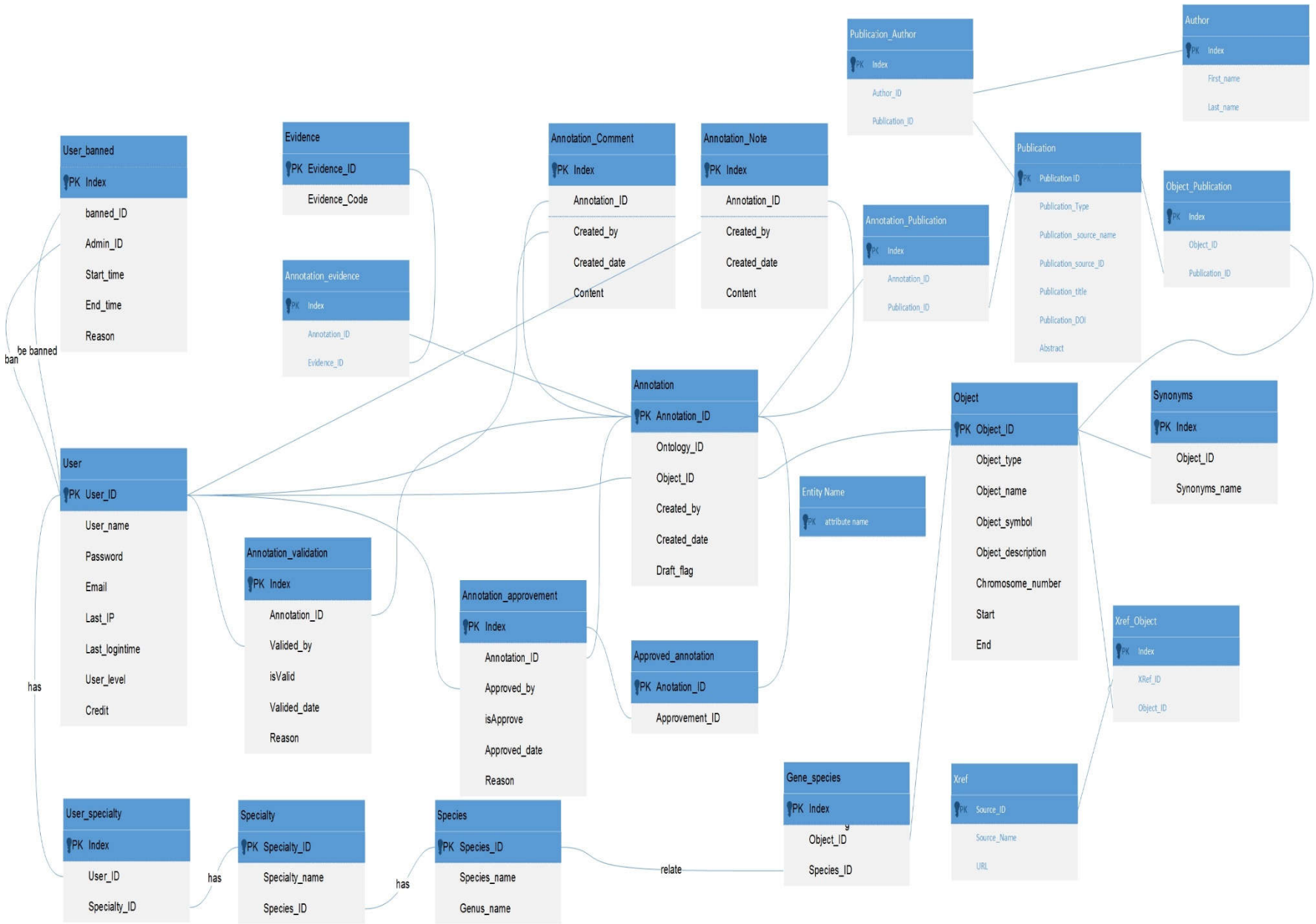


Figure 3: This diagram is a basic ER diagram of our web portal. Where we show the relationships (diamond shaped) between entities (box shaped). One can read a relationship like this, the entity gene has to have one species and that species is mandatory. Another example could be that the entity gene has many annotations or no annotations, so we call that optional to many.

6.2 Tables design

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


User_banned
PK Index
banned_ID
Admin_ID
Start_time
End_time
Reason
be banned
ban
Evidence
PK Evidence_ID
Evidence_Code
Annotation_evidence
PK Index
Annotation_ID
Evidence_ID
Annotation_Comment
PK Index
Annotation_ID
Created_by
Created_date
Content
Annotation_Note
PK Index
Annotation_ID
Created_by
Created_date
Content
Publication_Author
PK Index
Author_ID
Publication_ID
Author
PK Index
First_name
Last_name
Annotation_Publication
PK Index
Annotation_ID
Publication_ID
Publication
PK Publication ID
Publication_Type
Publication _source_name
Publication_source_ID
Publication_title
Publication_DOI
Abstract
Object_Publication
PK Index
Object_ID
Publication_ID
Annotation
PK Annotation_ID
Ontology_ID
Object_ID
Created_by
Created_date
Draft_flag
User
PK User_ID
User_name
Password
Email
Last_IP
Last_logintime
User_level
Credit
has
Entity Name
PK attribute name
Object
PK Object_ID
Object_type
Object_name
Object_symbol
Object_description
Chromosome_number
Start
End
Synonyms
PK Index
Object_ID
Synonyms_name
Annotation_validation
PK Index
Annotation_ID
Valided_by
isValid
Valided_date
Reason
Annotation_approvement
PK Index
Annotation_ID
Approved_by
isApprove
Approved_date
Reason
Approved_annotation
PK Anotation_ID
Approvement_ID
Xref_Object
PK Index
XRef_ID
Object_ID
User_specialty
PK Index
User_ID
Specialty_ID
has
Specialty
PK Specialty_ID
Specialty_name
Species_ID
has
Species
PK Species_ID
Species_name
Genus_name
relate
Gene_species
PK Index
Object_ID
Species_ID
Xref
PK Source_ID
Source_Name
URL

6.2.1 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		

6.2.2 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		

6.2.3 Table: Specialty

ROW NAME	DESCRIPTION	DATA TYPE	NULLABLE

Specialty_ID	Specialty ID PK		
Specialty_name	Specialty Name		
Species_ID	Species ID		

6.2.4 Table: User_Specialty

ROW NAME	DESCRIPTION	DATA TYPE	NULLABLE
Index	PK		
User_id	User ID		
Specialty_id	Specialty ID		

6.2.5 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		

6.2.6 Table: Synonyms

ROW NAME	DESCRIPTION	DATA TYPE	NULLABLE
Index	PK		
Object_ID	Unique Gene ID		
Synonym_name	Synonym Name		

6.2.7 Table: Species

ROW NAME	DESCRIPTION	DATA TYPE	NULLABLE
Species_id	Species ID, PK		
Species_name	Species Name		
Genus_name	Genus_name		

6.2.8 Table: Gene_Species

ROW NAME	DESCRIPTION	DATA TYPE	NULLABLE
Index	PK		
Object_ID	Gene ID		
Species_id	Species ID		

6.2.9 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		

6.2.10 Table: Annotation_Comment

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			

6.2.11 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			

6.2.12 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		

6.2.13 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		

6.2.14 Table: Approved_Annotations

ROW NAME	DESCRIPTION	DATA TYPE	NULLABLE
Index	PK		
Annotation_ID	Annotation ID		
Approvement_ID			

6.2.15 Table: Evidence

ROW NAME	DESCRIPTION	DATA TYPE	NULLABLE
Evidence_ID	Evidence ID, PK		
Evidence_Code	Evidence Code		

6.2.16 Table: Annotation_Evidence

ROW NAME	DESCRIPTION	DATA TYPE	NULLABLE
Index	PK		
Annotation_ID	Annotation ID		
Evidenc_ID	Evidence ID		

6.2.17 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		

6.2.18 Table: Annotation_Publication

ROW NAME	DESCRIPTION	DATA TYPE	NULLABLE
Index	PK		
Publication_ID			
Annotation _ID			

6.2.19 Table: Object_Publication

ROW NAME	DESCRIPTION	DATA TYPE	NULLABLE
Index	PK		
Publication_ID			
Object _ID			

6.2.20 Table: Author

ROW NAME	DESCRIPTION	DATA TYPE	NULLABLE

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

6.2.21 Table: Author_Publication

ROW NAME	DESCRIPTION	DATA TYPE	NULLABLE
index	PK		
Author_ID	Author ID		
Publication_ID	Publication ID		

6.2.22 Table: Xref

ROW NAME	DESCRIPTION	DATA TYPE	NULLABLE
Source_name	Source Name		
Source_ ID	PK, Unique ID for Xref_Object in the source		
URL	Reference to Ontology ACC		

6.2.23 Table: Xref_Object

ROW NAME	DESCRIPTION	DATA TYPE	NULLABLE
index	PK		
XRef_ID	XRef ID		
Object_ID	Object_ID		

7 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
[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
[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

[19] Varshney, Rajeev K., Ryohei Terauchi, and Susan R. McCouch. "Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding." (2014): e1001883.
[20] Varshney RK, Graner A, Sorrells ME (2005) Genomics-assisted breeding for crop improvement. Trends Plant Sci. 10: 621–630 [PubMed]
[21] https://en.wikipedia.org/wiki/Biocurator