Gene Curation Tool (GCT)

Design Document

08/24/2015

Contents

1 Introduction

- This long term project need a document to help participants to understand the background, terms, purpose, etc., So that they could comprehend the situation as soon as possible
- The developers and researchers, no matter come from computer science or biology, need a document to confirm the definition and verify the understanding of the whole project.
- This document will be the most authoritative handbook and reference while developing the project.

- The structure of this document and guide of reading
- Some foreword about this section, like purpose (intent to give enough definition and examples to help understanding)

1.1 Planteome Project

Why Planteome project is important?

What problem we try to solve by this project?

Basic information about 'Common Reference Ontologies and Applications (cROP) for Plant Biology'

1.2 Gene Curation Tool (GCT)

What is gene curation?

What the purpose of developing GCT?

Who will be the beneficiary from GCT and how? (the meaning of developing GCT)

1.3 Basic knowledge about Genes

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

Gene

DNA

RNA

Chromosome

Genome

Germplasm

QTL

Gene product

Phenotype

1.4 Ontology

What is ontology

Ontology in biology

Gene Ontology (GO)

What is GO? Some examples of GO

Plant Ontology (PO)

What is PO? Some examples of PO

Plant Trait Ontology (TO)

What is TO? Some examples of TO

Plant Environment Ontology (EO)

What is EO? Some examples of EO

Plant Stress Ontology (PSO)

What is PSO? Some examples of PSO

1.5 Annotation

What is Annotation?

What's the use of Annotation?

Where the Annotation data come from?

GO Annotation examples

2 Existing Ontology Databases

2.1 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.

2.1.1 Introduction

What is Amigo?

2.1.2 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

2.1.3 GO Annotation File (GAF) Format

Introduction of Annotation File Fields (GAF 2.1)

The explanation of the format

2.2 Gramene

Introduction:

Features:

2.3 AgriGO

Introduction:

Features:

2.4 Conclusion

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

3 Software Requirement Specification

Itemize the user requirements,

3.1 Product Perspective

Who would need to use GCT?

The benefit of using GCT

The meaning of using GCT

3.2 Scope

We descript the features in scope of GCT.

a. …
b. …

3.3 Operating Environment

Database: MySQL

Server: Apache

3.4 Role Based Access Control (RBAC)

What is RBAC?

How we use it in our system?

3.5 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.)	

3.6 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					

3.7 Contribution Credit Scheme

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**	

3.8 User management subsystem

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

3.8.1 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	

3.8.2 Use Case: Login
3.8.3 Use Case: Ban User

…

3.8.4 Use Case: Edit Specialty

…

3.9 Annotation management subsystem
3.9.1 Use Case: browse/edit/add annotation
3.9.2 Use Case: Save annotation draft
3.9.3 Use Case: Save note
3.9.4 Use Case: flag annotation
3.9.5 Use Case: Comment on annotation
3.10 Object management subsystem

…

3.11 Publication management subsystem

…

4 API design

In this section, all the APIs related to Gene Curation Tool system will be descripted.

4.1 Object import API

We need to develop an API to import the Object data

4.2 Annotation import API

API to import the annotation

4.3 Annotation export API

API to export the annotation

4.4 Utilize API to get Ontology information

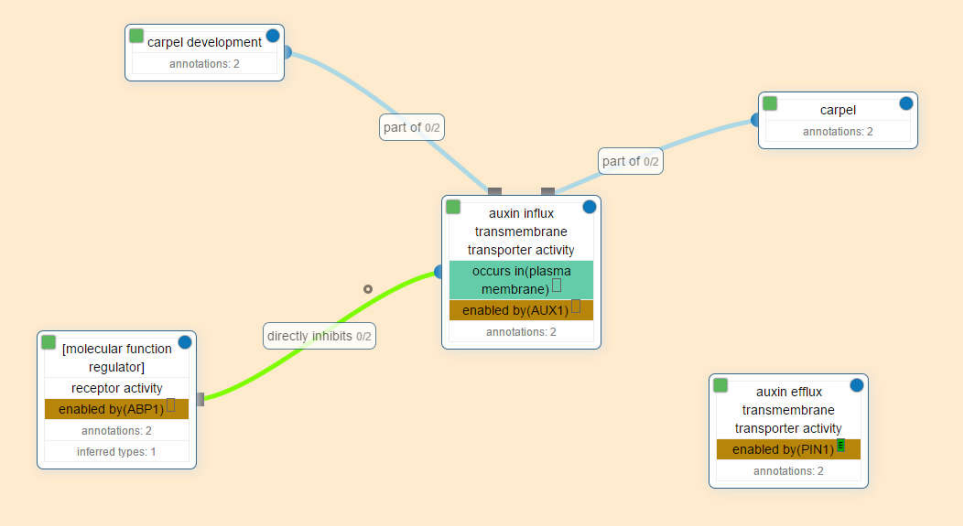
The ontology data should be accessed by using Amigo API

5 User Interface Design

The hand drawing user interface design

5.1 Web Pages Design

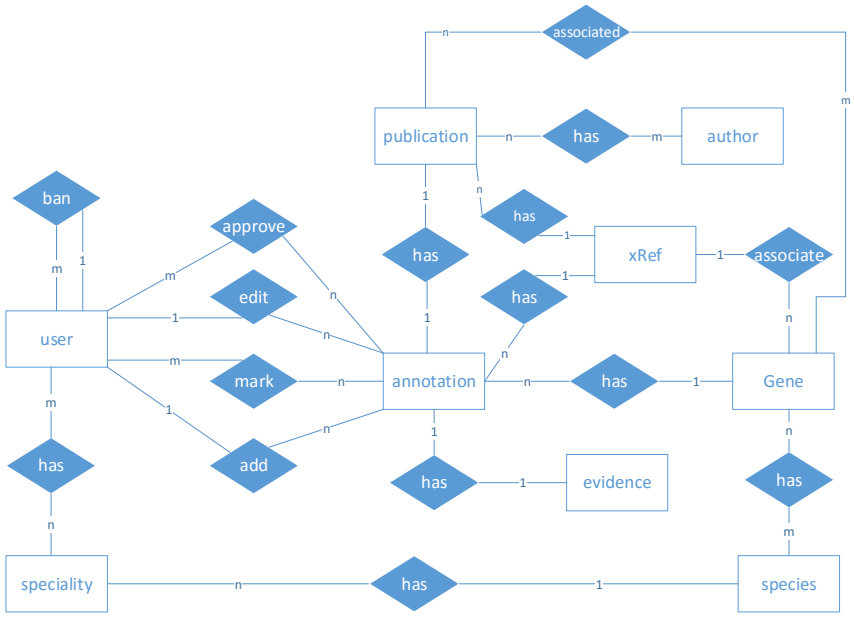
5.2 Graphic Gene Curation



6 Database Design

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

6.1 ER diagram



6.2 Tables design

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

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		

[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] …

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, or macromolecules, consisting of one or more long chains of amino acid residues” [8].