Araport Chado Extension Data Model

Reference
Version 1.0

February 2015

Revision History

Date	Version	Description	Author
February 3, 2015	1.0	Document created	Belyaeva I

Table of Contents

Introduction

Conceptual Overview

The proposed approach involves several Chado Modules, including Sequence, Genetic, Phenotype, Publications, and Stock Module.
The methodology is similar how TAIR associates locus to phenotypes with exception that Chado schema stores data using different entities arrangement.
The prerequisites to make associations between locus, phenotypes, and stocks is to load baseline reference data:

- Gene Data
- Publications Data
- Phenotype Data
- Stock Data
- Allele Data
- Genotype (Polymorphism)

There are several steps to make cross-reference associations work:

Locus -> Allele -> Phenotype Path

  - Locus and Allele Association
  - Allele and Phenotype Association

Additionally, we have to make associations between Stock data and Genotype (enclosed entity for allele sequrcne) to ensure Chado/Tripal modules function correctly.

- Stock - > Genotype
  - Association of Genotype and Allele
  - Association of Stock and Genotype
- Stock - > Phenotype
  - Association of Stock and Phenotype

Evidence Association Path:

- Phenotype to Publication
- Stock to Publication

TAIR uses direct association of Polymorphism data to locus (it stores SNP in the same entity as polymorphism itself). However, the rest of associations are similar.

TAIR Locus to Phenotype path:

- Locus to Polymorphism
- Locus to Germplasm
- Germplasm to Phenotype

The figures below graphically depicts the conceptual data flow between key entities and their relationships.

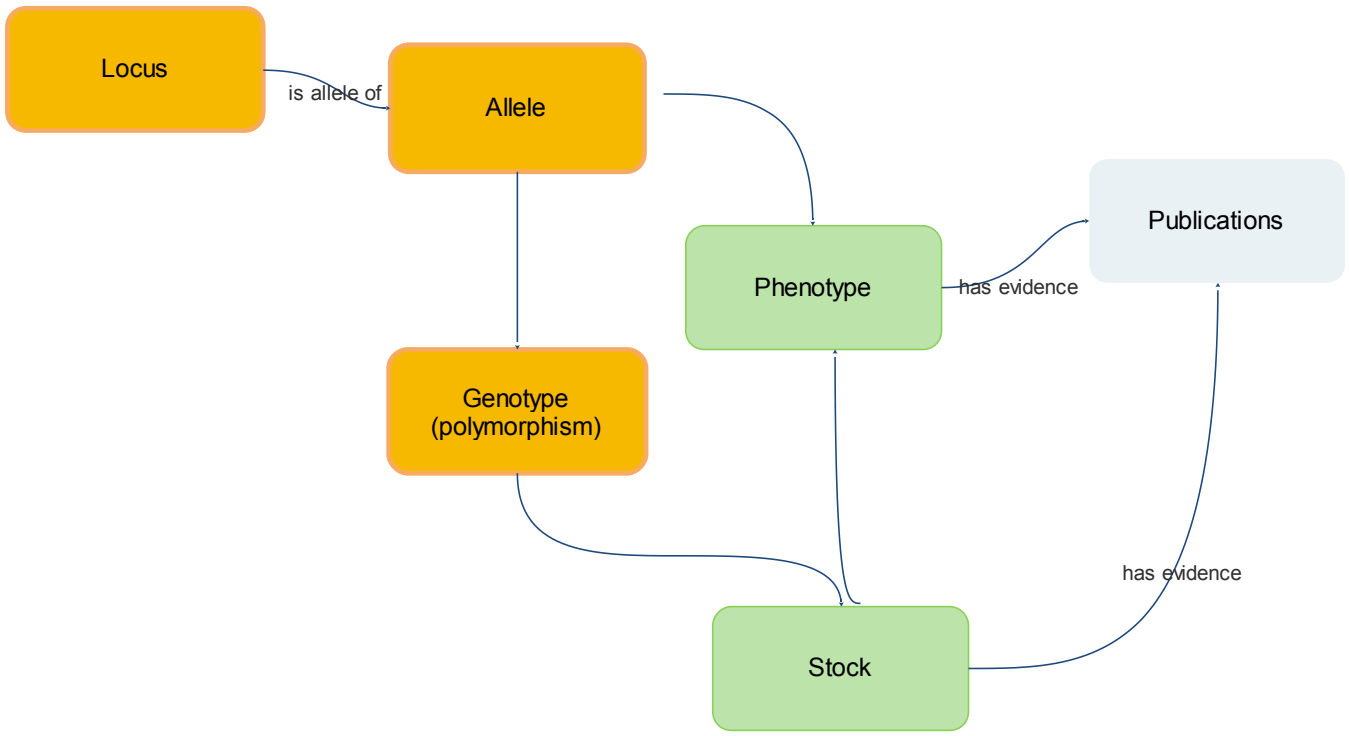


Figure 1.1 Schematic path of Locus, Allele, Phenotype and Stock Associations.

Association of Locus, Allele and Phenotypes

Approach Overview

The **locus data** is stored in the **Sequence module** using **feature** entity with locus data annotated as **gene** type. The allele data is stored in the **Sequence module** using **feature** entity with locus data, however, annotated as **sequence** type.

Locus -> Allele -> Phenotype Path

The allele sequence data association is recorded using **feature_phenotype** entity, where feature id attribute is the **allele feature id** and phenotype id attribute is the referenced phenotype.

Finally, **locus feature** is associated with **allele feature** using **feature_relationship** entity with annotated relationship type is **allele of**.
The locus plays role of a parent, and allele sequence is a child.
Subject id: locus feature Id
Object Id: allele feature Id.
Relationship Type: **is allele of**.

The link of associations, from locus to allele, from allele to phenotype form a granular association path between **locus** and **phenotypes**.

The diagram on Fig 1.2 depicts the nature of associations and navigation paths between entities.

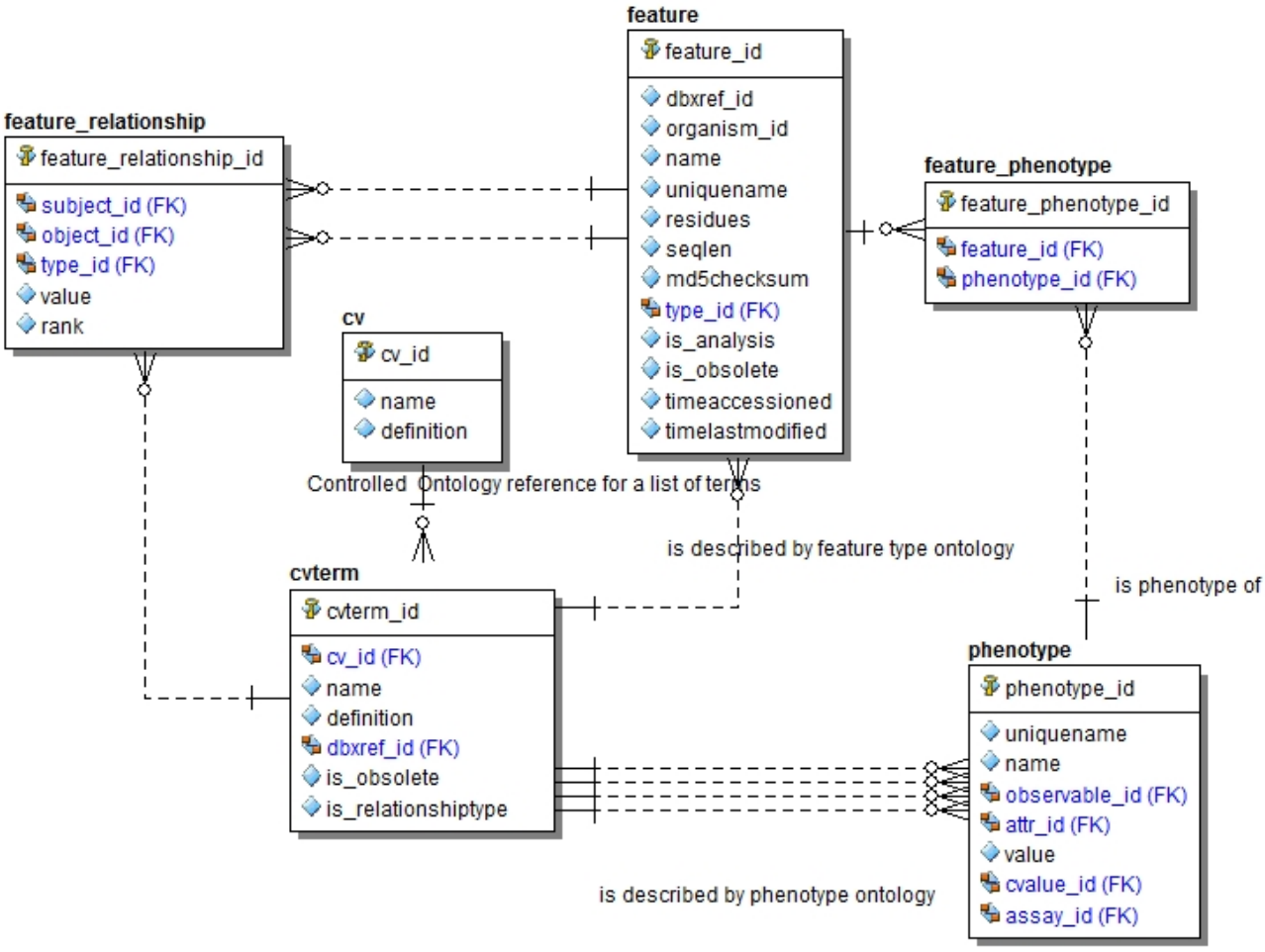


Figure 1.2 Locus, Allele and Phenotype Associations.

Stocks and Phenotypes Association

Approach Overview

The proposed approach is a three-step approach, which includes:

- Association of genotype and allele sequence
- Association of stock and genotype data
- Association of stock and phenotype data.

The last step can be done using natural diversity module or create an entity, which would capture association between stocks and phenotypes directly. Tomato Chado uses natural diversity module since the phenotyped data was collected as a result of real natural diversity experiments.

The original contributor of the Chado Stock Module, ParamediumDb suggested to create the linkage entity **stock_phenotype**, and therefore bypass the complexity of stocks to phenotype linkage incurred by Natural Diversity Module.

- Association of genotype and allele sequence

The genotype is associated with its set of alleles using **feature_genotype** entity, where feature id attribute is the **allele feature id** and genotype id attribute is the referenced genotype.

- Association of stock and genotype data

The stock and genotype (stock's set of polymorphisms) is associated using **stock_genotype** entity, where **stock_id** attribute is the stock with associated polymorphisms and **genotype id** attribute is the referenced genotype.

- Association of stock and phenotype data

Using **stock_phenotype** entity

The stock and phenotype (stock's set of phenotypes) is associated using **stock_phenotype** entity, where **stock_id** attribute is the stock with associated phenotypes and **phenotype id** attribute is the referenced phenotype.

Using **Natural Diversity Module**

Each Stock is stored in ND_EXPERIMENT_STOCK entity.
Each Phenotype is stored in ND_EXPERIMENT_PHENOTYPE entity.

Each Genotype is stored in ND_EXPERIMENT_GENOTYPE entity.
Finally, by using common value for ND_EXPRIMENT_ID we can associate Stock and related phenotypes and genotypes.

This type of association is depicted on Fig 1.4.

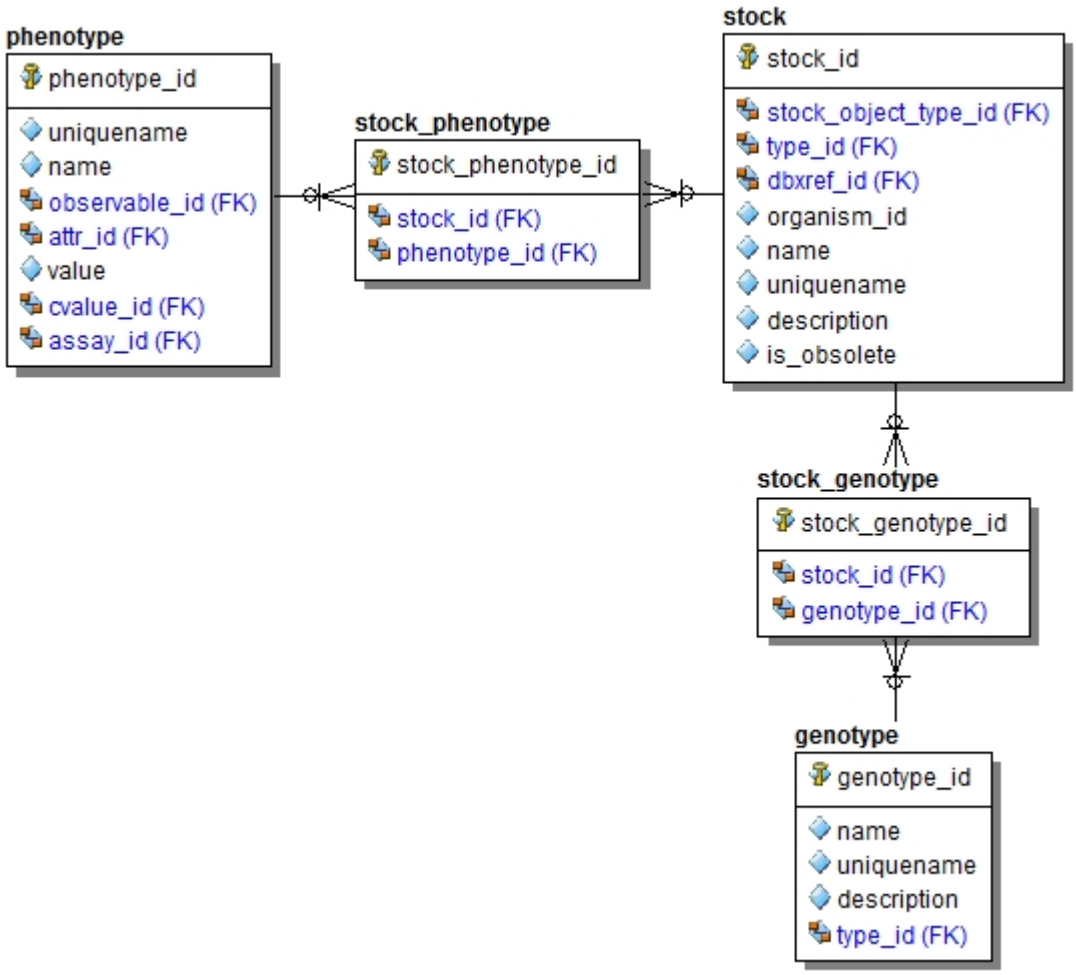


Figure 1.3 Stock, Phenotypes and Genotype Associations.

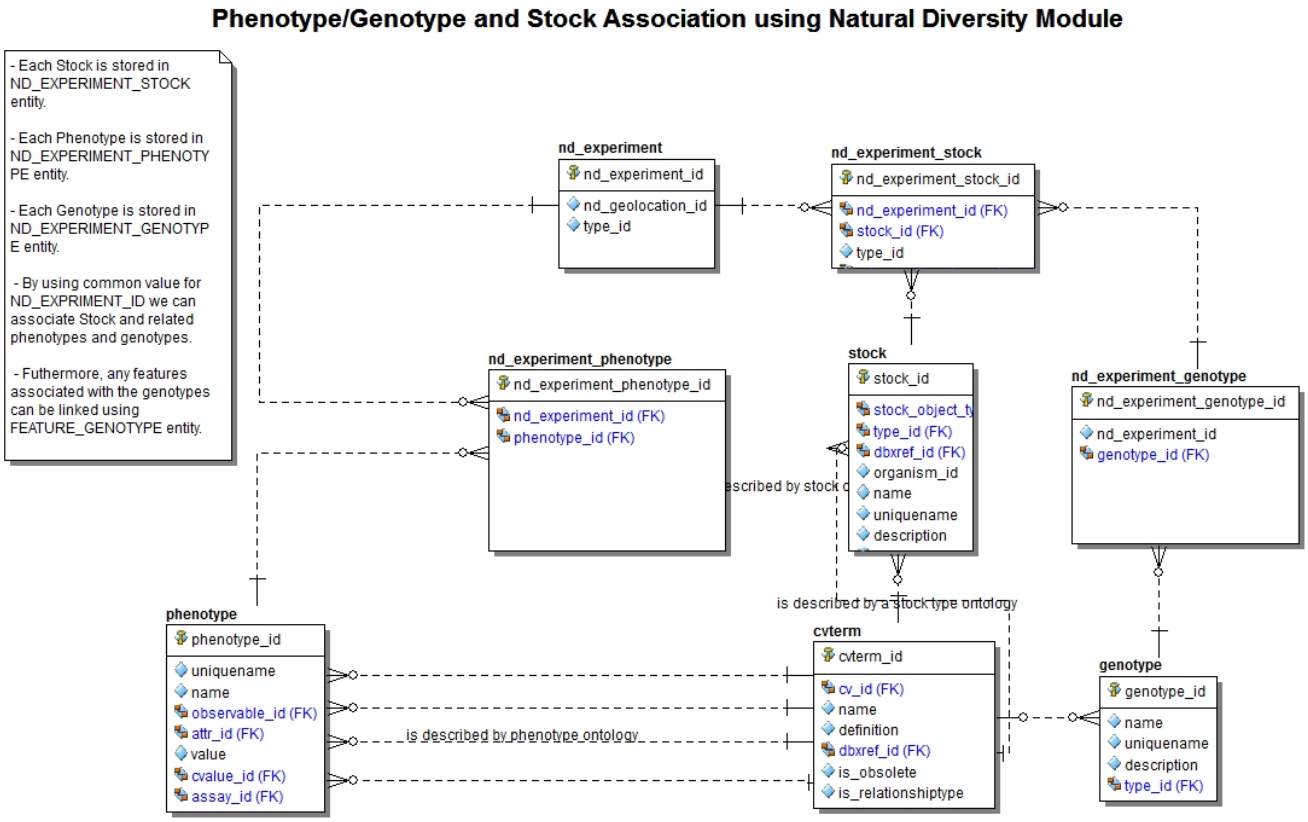


Figure 1.4 Stock, Phenotypes and Genotype Associations using Natural Diversity Module

Association of Stock, Publications and Phenotypes

Approach Overview

- Matching and Loading Publications

First, since Araport Chado integrated database has a set of preloaded publications associated with PubMed Id we would need to match loaded Chado publications with TAIR publications as source of evidence for stocks and phenotypes.

If Araport Chado database does not have matching publication, which exists in TAIR database, we would need to load publication data from TAIR.

- Linking Stocks and Publications

Stock entity instance is associated with publication using **stock_pub** entity. The **stock id** attribute is a foreign key to the stock entity, and **pub_id** is the referenced **publication_id.**

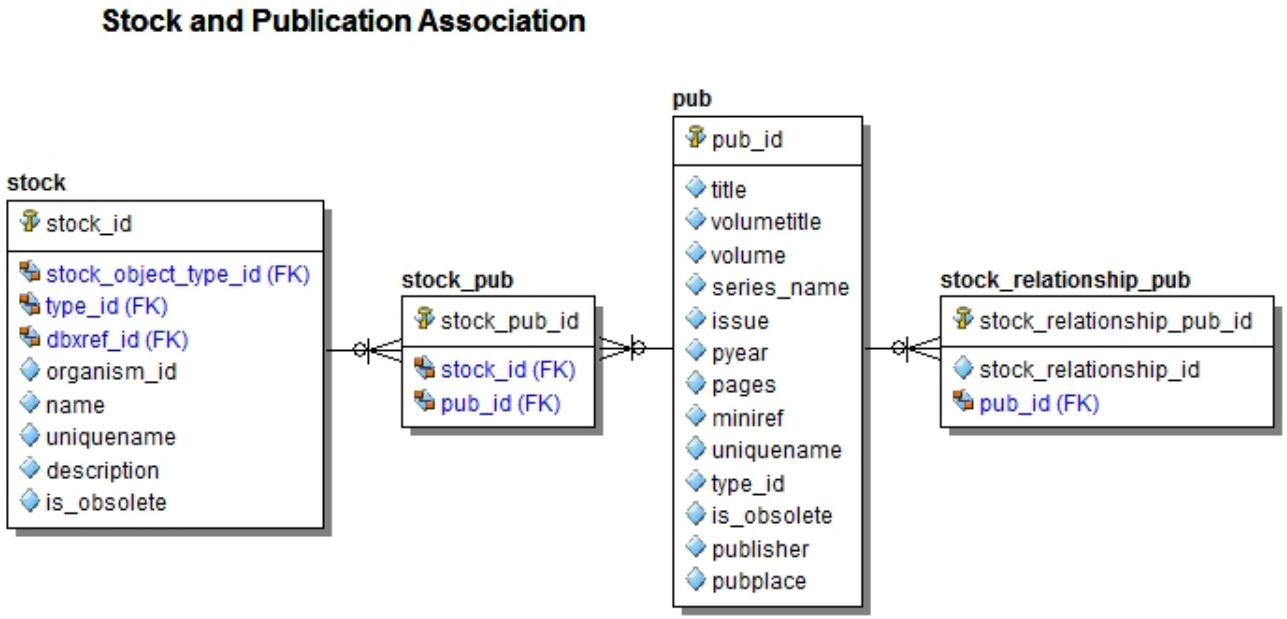


Figure 1.5 Stock and Publication Associations.

- Linking Phenotypes and Publications

Phenotype and publications are linked using **phenstatement**, and **phendes**c entities. **Phendesc** is the description of the phenotype per associated genotype, environment, and publication. Phenstatement is more granular representation of a phenotype with ability to resolve phenotype description per each associated publication.

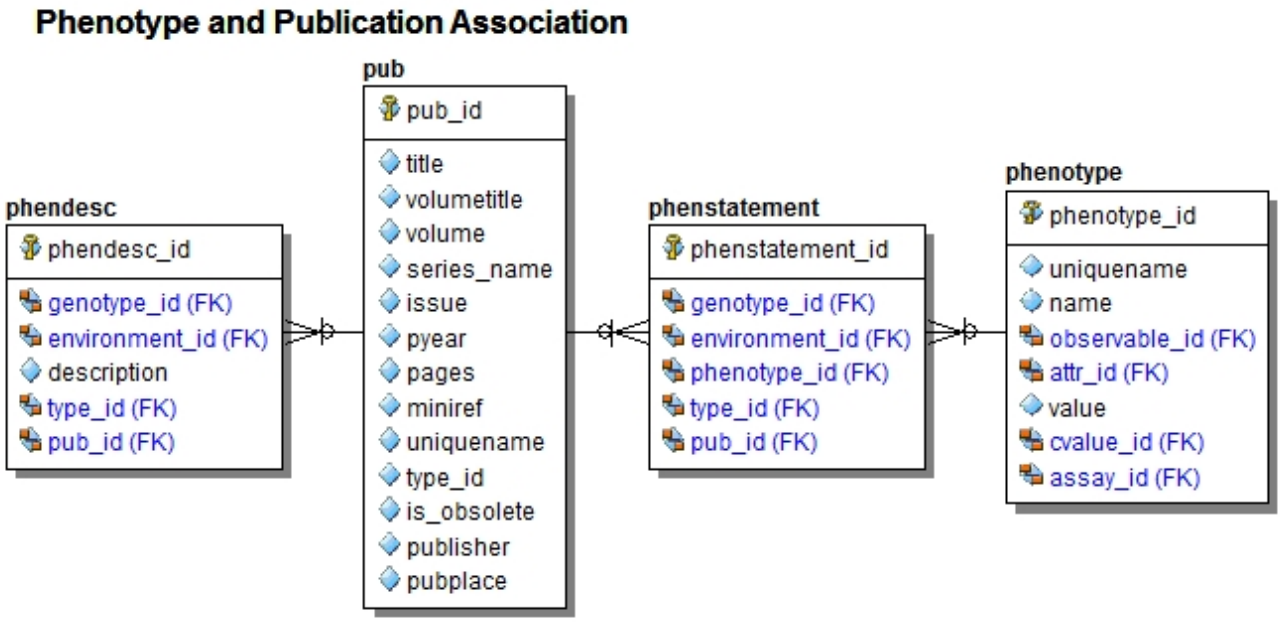


Figure 1.6 Phenotype and Publication Associations.

Stock Images

Approach Overview

The Chado schema is extended to store images, their metadata, and lightweight preview.
Stock records can be associated with multiple images using linking table **stock_image**. Stock image is assigned a rank attribute to denote an order of the image display. The stock image ERD is depicted on Fig.1.7 "**Chado Stock Images Association**".

Participating Entities:

- md_image – image metadata
- md_image – image data
- Md_image – image preview
- stock_image – list of stocks and associated entities

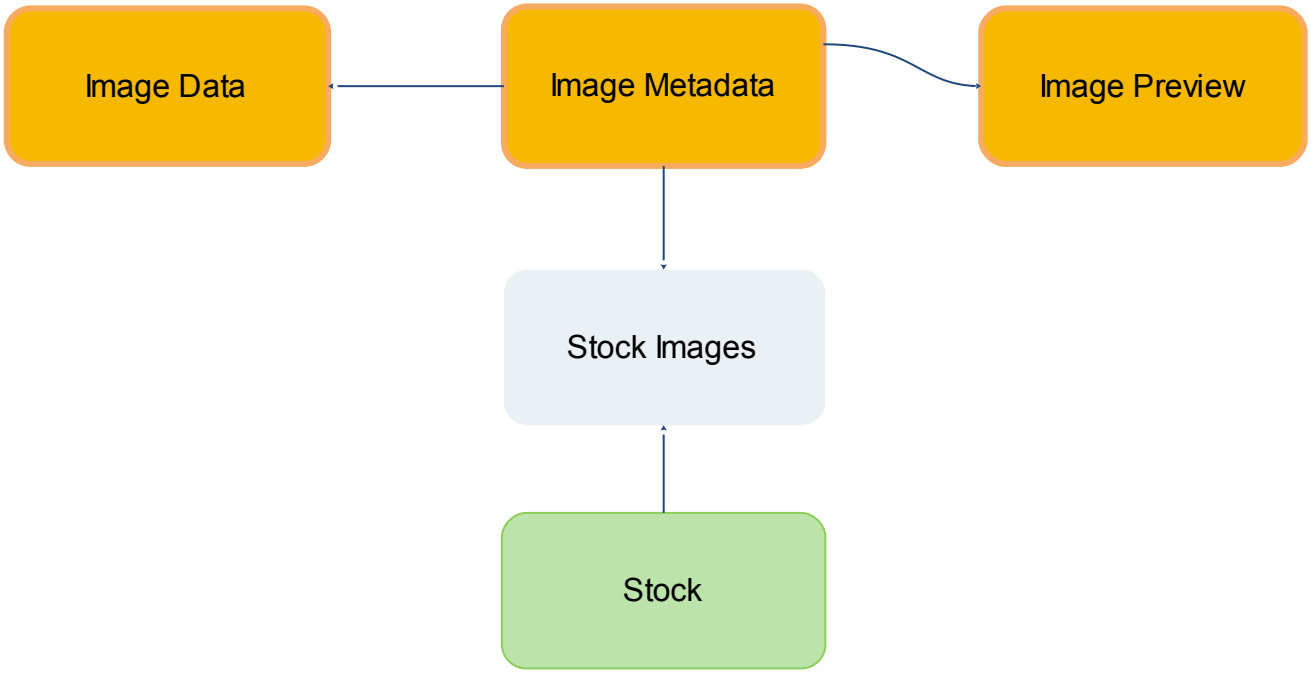


Figure 1.7 Chado Stock Images Conceptual View

The TAIR schema has a similar conceptual data layout for storing stock images with exception of the schema specific naming and design conventions (**Fig 1.9 TAIR Image Metadata**).

Stock Image Entity Cluster Data Dictionary

#	Entity Name	Description
1	md_image	Image metadata. Data store for essential image attributes such as name, description, image type, and file extension. Image type – scanned image, photograph. Image type is a cvterm of image_type. File –format –(gpeg, gif, etc.)
2	md_imagedata	Image data. Data store for actual image data in a binary format.
3	md_imagethumbnail	Image data. Data store for image preview data in a binary format.
4	stock_image	Images associated with a stock. Image have a rank attribute for display purposes.
5	stock	Data store for stocks. Referenced in the Stock Image Entity cluster.

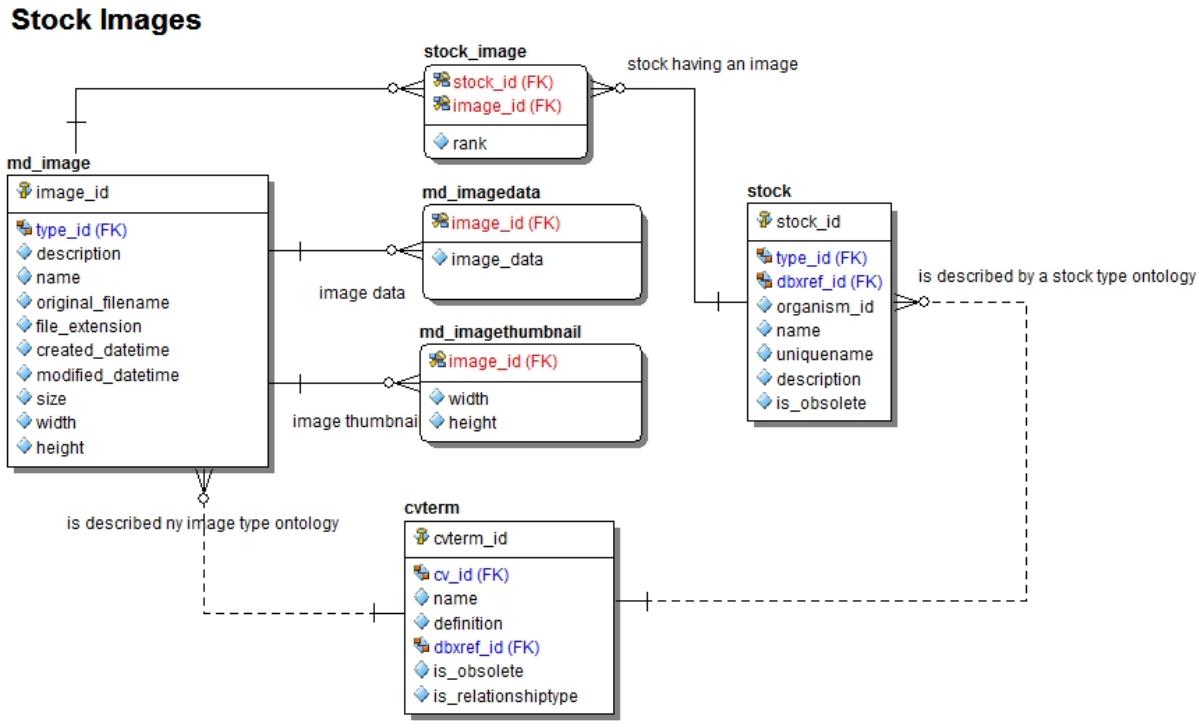


Figure 1.8 Chado Stock Images Association

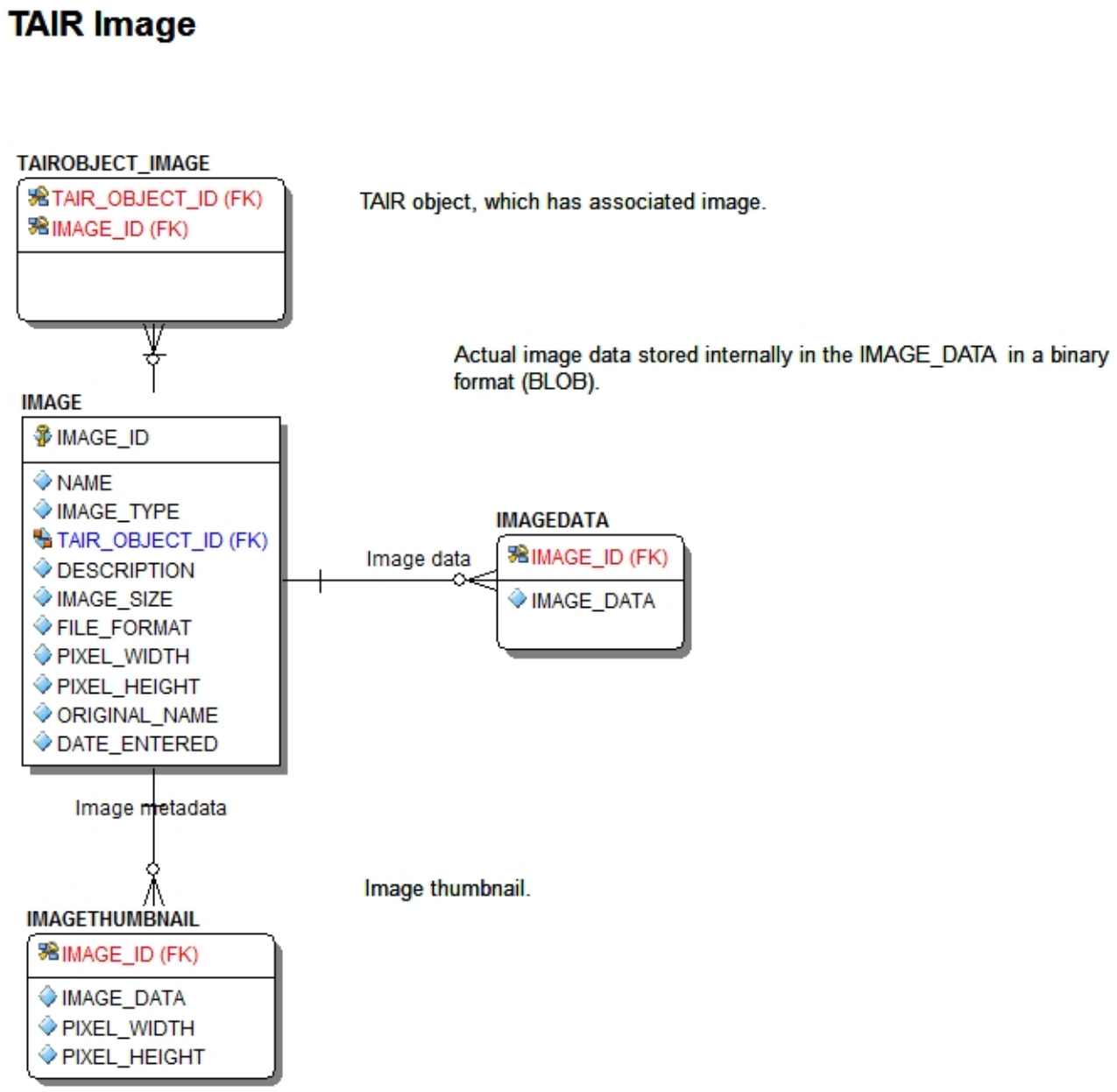
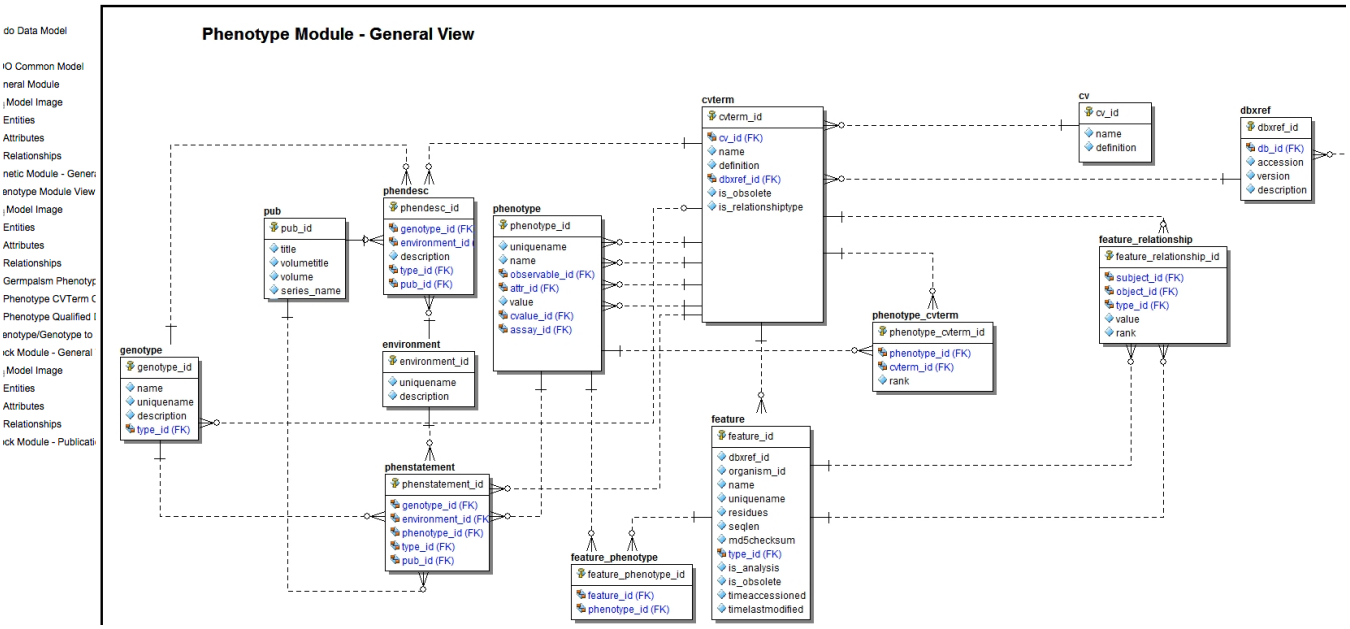


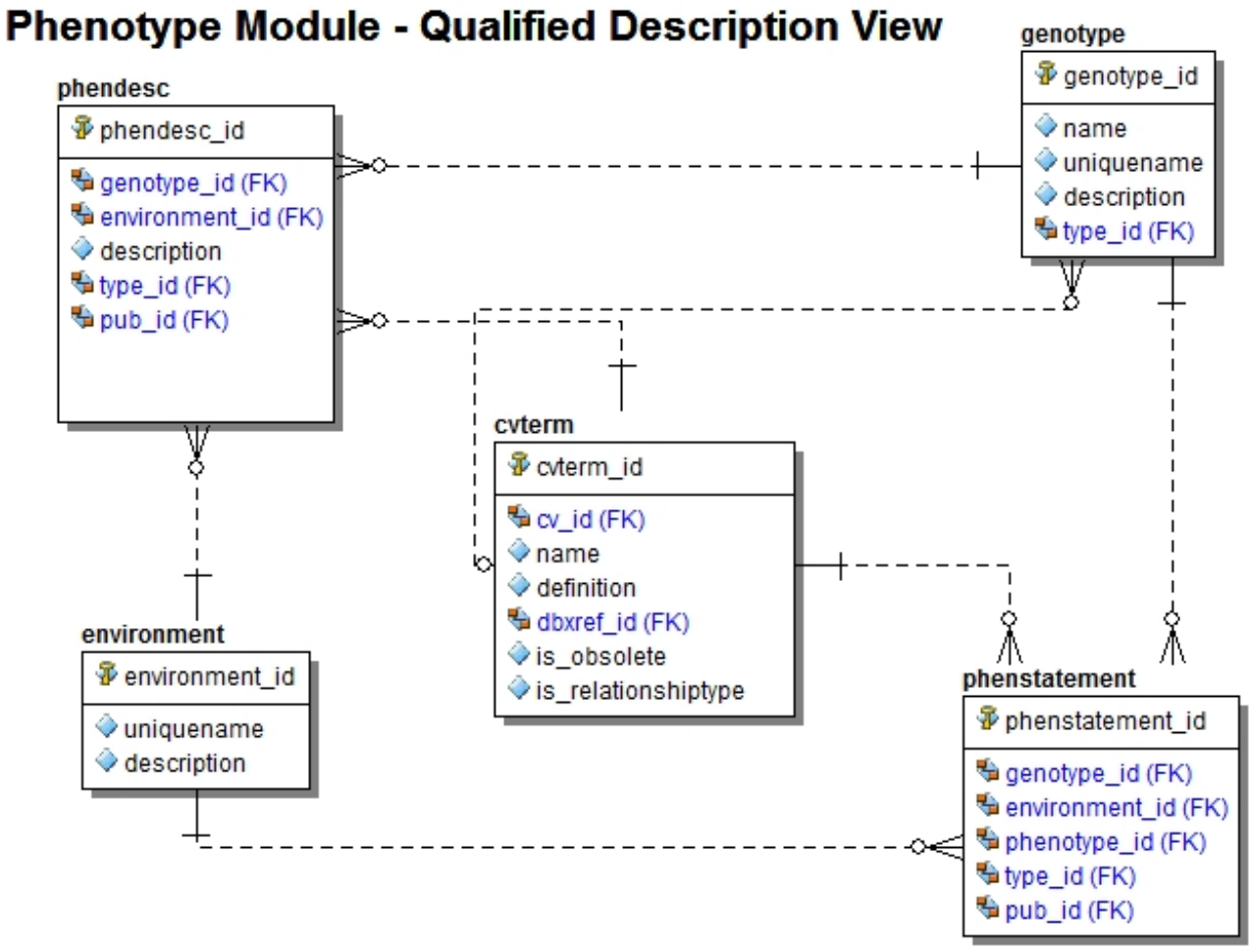
Figure 1.9 TAIR Image Metadata

Appendix

Phenotype Module Entity-Relationship Diagram

Araport Chado Stock Module Data Model Web Report




Phenotype Module - Qualified Description View
phendesc
phendesc_id
genotype_id (FK)
environment_id (FK)
description
type_id (FK)
pub_id (FK)
genotype
genotype_id
name
uniquename
description
type_id (FK)
cvterm
cvterm_id
cv_id (FK)
name
definition
dbxref_id (FK)
is_obsolete
is_relationshiptype
environment
environment_id
uniquename
description
phenstatement
phenstatement_id
genotype_id (FK)
environment_id (FK)
phenotype_id (FK)
type_id (FK)
pub_id (FK)

References

1. Jung, S., Menda, N., Redmond, S., Buels, R. M., Friesen, M., Bendana, Y., ... & Main, D. (2011). The Chado Natural Diversity module: a new generic database schema for large-scale phenotyping and genotyping data. *Database*, *2011*, bar051.