

---

# Report on Task 22 - Development of Generation Challenge Programme Domain Models

**Richard Bruskiewich<sup>1</sup>, Guy Davenport<sup>2</sup>, Tom Hazekamp<sup>3</sup>,  
Thomas Metz<sup>1</sup>, Manuel Ruiz<sup>4</sup>, Reinhard Simon<sup>5</sup>, Masaru  
Takeya<sup>6</sup>, Jennifer Lee (U. Dundee/SCRI)<sup>7</sup>, Martin Senger<sup>1</sup>,  
and Graham McLaren<sup>1</sup>**

---

<sup>1</sup>IRRI, <sup>2</sup>CIMMYT, <sup>3</sup>IPGRI, <sup>4</sup>CIRAD, <sup>5</sup>CIP, <sup>6</sup>NIAS, <sup>7</sup>University of  
Dundee/Scottish Crop Research Institute



---

# Report Overview

- Synoptic (whirlwind) tour of the EclipseUML models as they sit as of today(in Cropforge)
  - Package structure
  - Model highlights
  - Comments on model design, especially, metadata models
- Future directions revisited...
- Proposed structure for next year's Task work?

# GCP Middleware Package Structure

- **Root package:**
  - org.generationcp
- **Demeter models:**
  - org.generationcp.model.germplasm.Germplasm
- **Ceres datasource-independent implementations of a model:**
  - org.generationcp.model.germplasm.**impl**.Germplasm**Impl**
- **Osiris datasource-specific implementations of the model:**
  - org.generationcp.**datasource**.db.hibernate.germplasm.Germplasm**DataImpl**

# Demeter Design Preamble

- Demeter model elements specified mostly with <<interface>> UML stereotype
  - Exception classes being the exception, of course...
  - Mostly one-to-one with Ceres implementations except for some multiply inherited interfaces.
- UML “best practices” lightly followed:
  - Standard “get/setter” classes especially EclipseUML generated ones not usually shown.
  - Primary (database) key identification of elements generally implicitly assumed but not modeled (except for “Entity”)
- Curious EclipseUML artifacts occasionally lurk...

# Top Level Demeter Packages

## Core Metadata Models:

- org.generationcp.model.metadata

## General Purpose (Data) Entity Models:

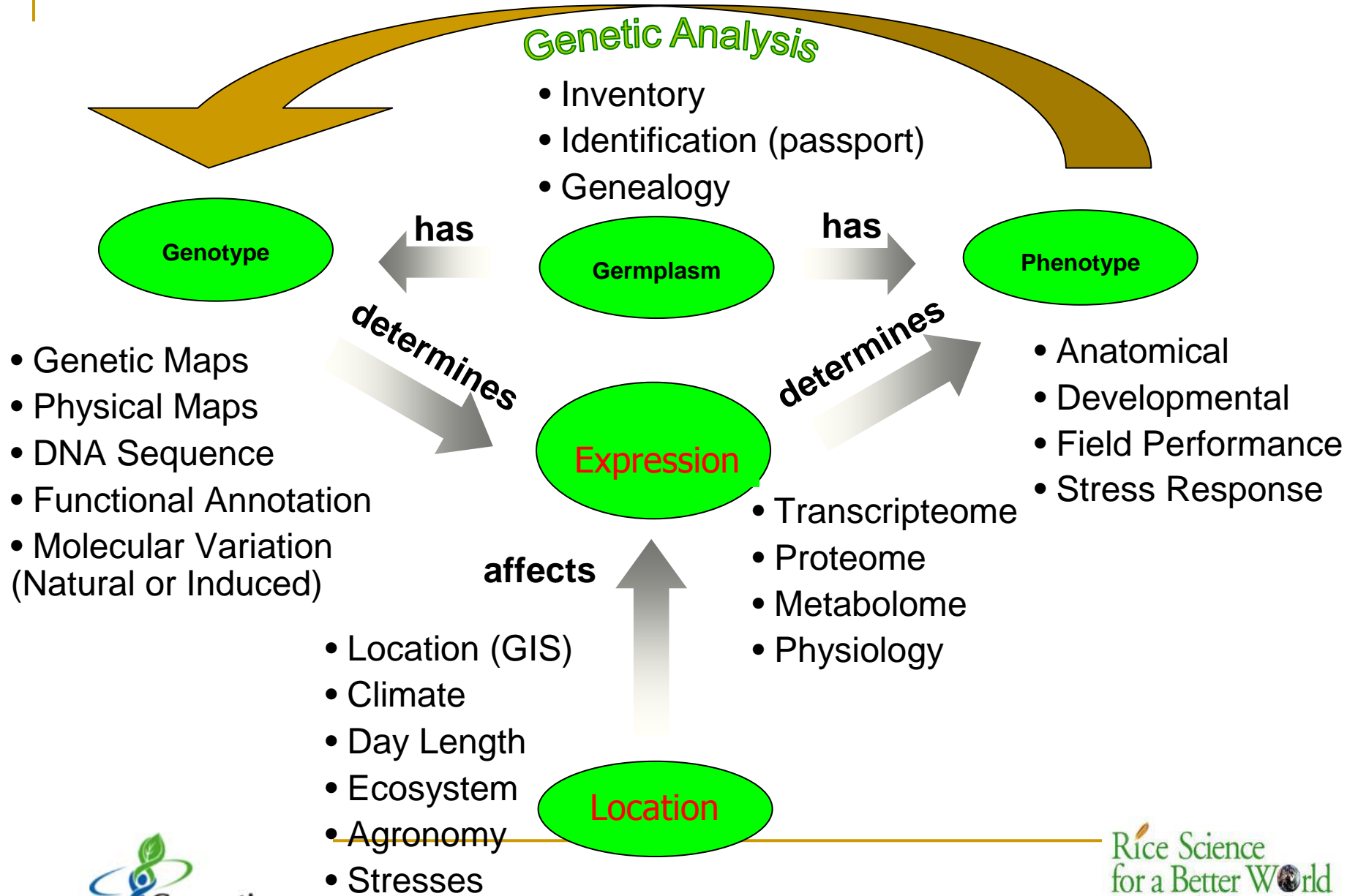
- org.generationcp.model.datasource
- org.generationcp.model.documentation
- org.generationcp.model.organization
- org.generationcp.model.study
- org.generationcp.model.util

## Scientific Entity Models:

- org.generationcp.model.expression
- org.generationcp.model.genotype
  - org.generationcp.model.genotype.map
- org.generationcp.model.germplasm
  - org.generationcp.model.germplasm.passport
- org.generationcp.model.location
- org.generationcp.model.phenotype



# Scientific Entity Model Scope



---

# Demeter Metadata Packages

## Root Package

- org.generationcp.model.metadata

## Feature Packages

- org.generationcp.model.metadata.feature
- org.generationcp.model.metadata.feature.identifier
- org.generationcp.model.metadata.feature.value
- org.generationcp.model.metadata.feature.evidence

## Identification Packages

- org.generationcp.model.metadata.namespace
- org.generationcp.model.metadata.namespace.nomenclature

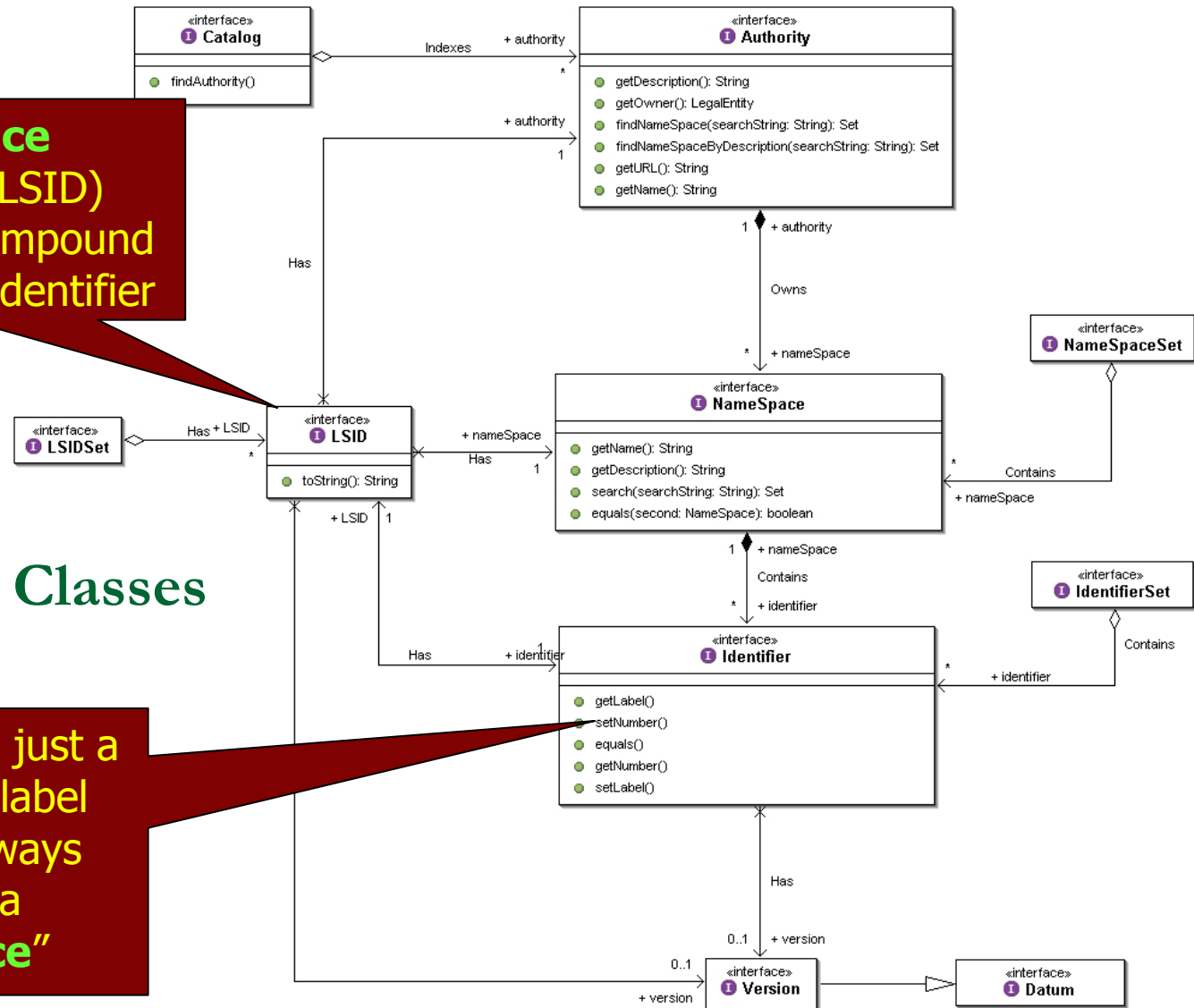
## Controlled Vocabulary & Ontology

- org.generationcp.model.metadata.ontology

**Life Science Identifiers (LSID) modeled as a compound (multi-layered) Identifier**

## Identification Classes

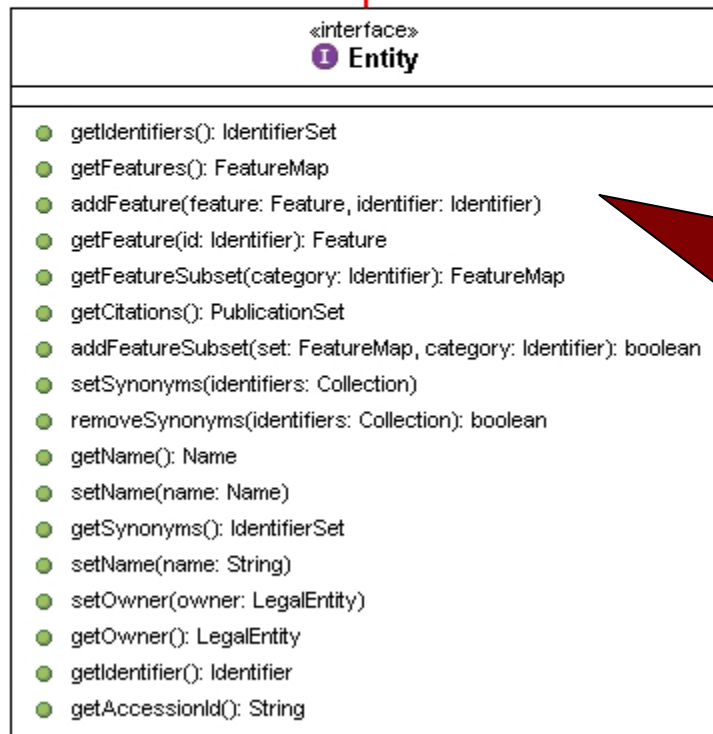
**An Identifier is just a String used to label something. Always belongs to a "NameSpace"**





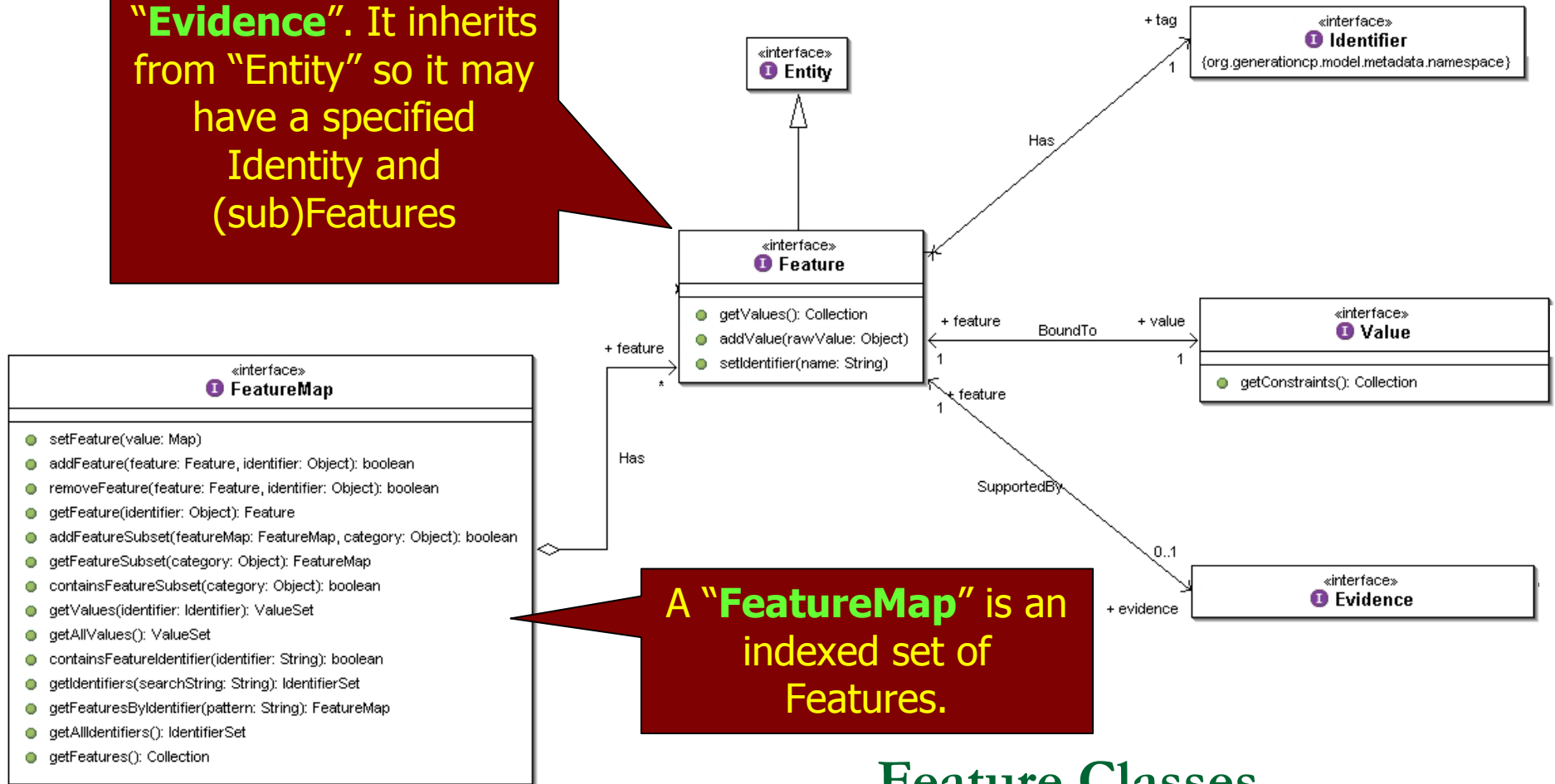
Every Entity has a primary identifier which is an AccessionId (which generally resolves to its primary key in the underlying database)

## Entity Class



An **Entity** is any "thing" (including abstract concepts like "location") that has distinct instances in the system that we wish to track, that have **Identity** and **Features**

A **Feature** is an **"Identifier-Value"** assertion, possibly with **"Evidence"**. It inherits from **"Entity"** so it may have a specified Identity and (sub)Features

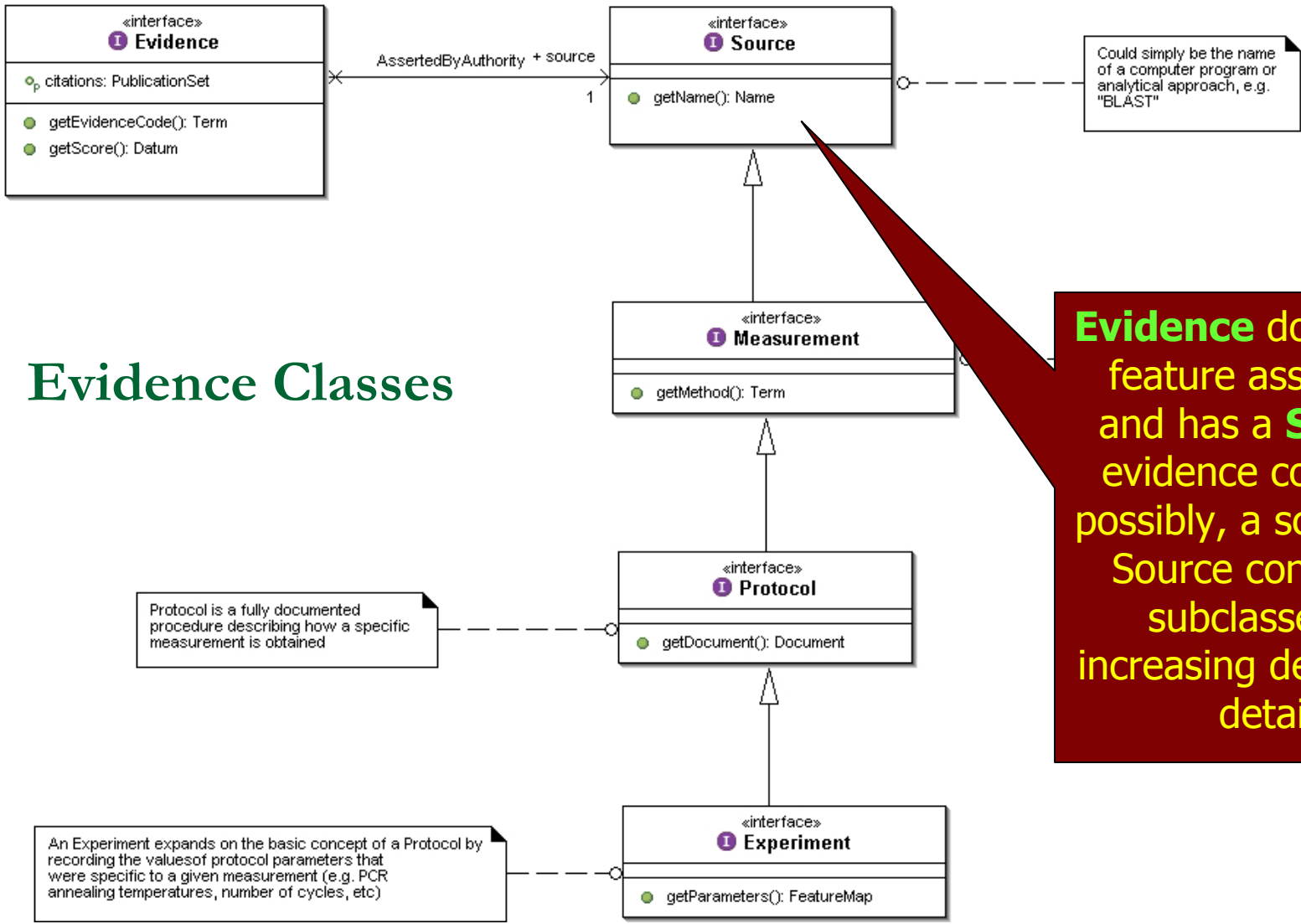


A **"FeatureMap"** is an indexed set of Features.

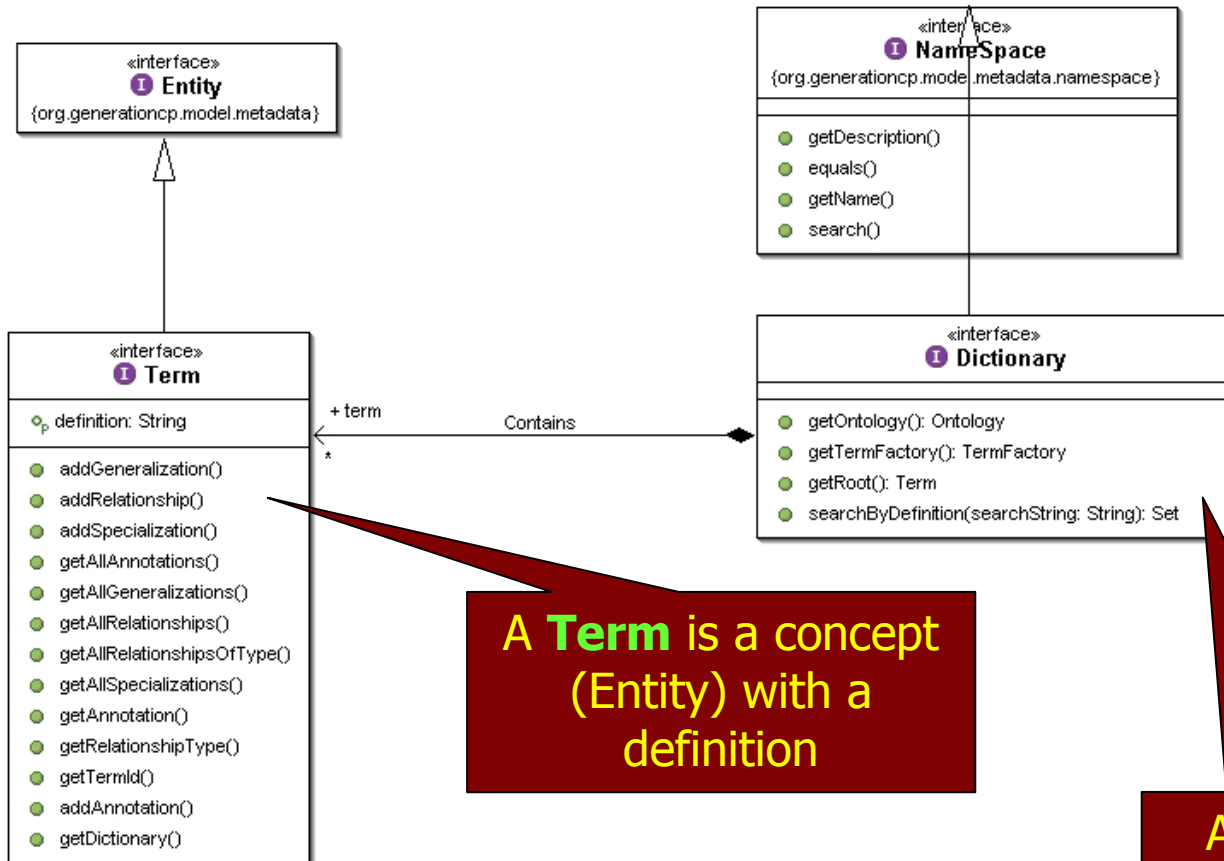
## Feature Classes



# Evidence Classes



**Evidence** documents feature assertions and has a **Source**, evidence code and possibly, a score. The **Source** concept is subclassed to increasing degrees of detail



A **Term** is a concept (Entity) with a definition

A **Dictionary** is a NameSpace (defined collection) of Terms

## Controlled Vocabulary & Ontology Classes

---

## More models...

- Too many big image files were presented in the Rome meeting at this point in the presentation; these image files will be reposted to the CropWiki for casual browsing, at:

[http://cropwiki.irri.org/gcp/index.php/EclipseUML\\_Consolidated\\_2005\\_Domain\\_Models](http://cropwiki.irri.org/gcp/index.php/EclipseUML_Consolidated_2005_Domain_Models)

- For the more technically inclined, the full (computable) model files are at CropForge (<http://cropforge.irri.org>) under the GCP Middleware “Demeter” subproject, as Omondo Eclipse UML class diagram files.

# Future Directions... Revisited

- Formal release of version 1.0 models coupled with a scientific paper submission about the project by year end
- So far, the modeling activity has been mostly an armchair analysis of domain entities and attributes... In the near future, more attention will be paid to the iterative validation of the domain models in GCP platform implementation of high priority use cases (such as the scenario just presented).
- Wish to develop a strategy to automatically generate data templates (essentially, XML schemata) directly from the domain model.
- Wish to (automatically?) map the domain models onto web service protocols (i.e. onto BioMOBY data types).
- **Q: How should the next generation of models be specified?**
  - **Using EMF (as proposed by Guy Davenport)**
  - **More usage of semantic web specifications (i.e. RDF/OWL specifications using Protégé?)**

# Proposed Structure for next year's work?

Split the task into two parts:

1. (\$80K) “Arm chair” refinement and extension of the existing domain models and ontology; Face-to-face meetings were very helpful for task for accelerated model development and integration this year, so plan to invite ~20 scientific domain experts and data curators (not just bioinformaticians) to a GCP domain model review and ontology curation/integration jamboree in mid-2006
  - Accelerate adoption/development of GCP ontology: for (drought) phenotyping, environment descriptors, germplasm genealogy, etc.
  - New public initiatives to engage: FUGO @ MGED
2. (\$120K) Iterative validation and refinement of existing domain models by (Java+data template+web service) platform implementation of the domain models by a dedicated core group of GCP software engineers, guided by priority use cases