A Semantic Wiki for Genotype/Phenotype Annotation in Plant Species

Justin Preece[1], Justin Elser[1], Pankaj Jaiswal[1]

[1]Department of Botany and Plant Pathology, 2082 Cordley Hall, Oregon State University, Corvallis, OR, 97331-2902, USA

Introduction

The number of sequenced plant genomes has grown in recent years and this growth underscores the necessity of providing quality gene annotations for biochemical, phenotype, expression and allelic variation characteristics. Except for a few model plant species, a majority of the sequenced genomes do not have a dedicated home on the Web to host this information. Therefore, the plant sciences research community needs more effective web annotation tools and semantic integration of annotations with their annotated data objects (*e.g.* genes).

Many current projects rely on a potpourri of desktop applications and static web forms to acquire annotations for research databases. There are also several web applications that encourage the curation of data using the MediaWiki platform (http://www.mediawiki.org), but most of these sites are either focused on individual genomes or a particular research sub-field (*i.e.* proteins, biological pathways, or SNP's only). Further, many of the extant biological wikis rely on free-form text blocks or simple iterative controls for the containment and structuring of curatorial data. The Plant Ontology Consortium (http://www.plantontology.org) proposes to build a semantic wiki site that simultaneously makes annotation more convenient, provides rigorous semantic data structure, includes a curatorial approval process, and offers publication and citation incentives to curators to participate in the annotation process.

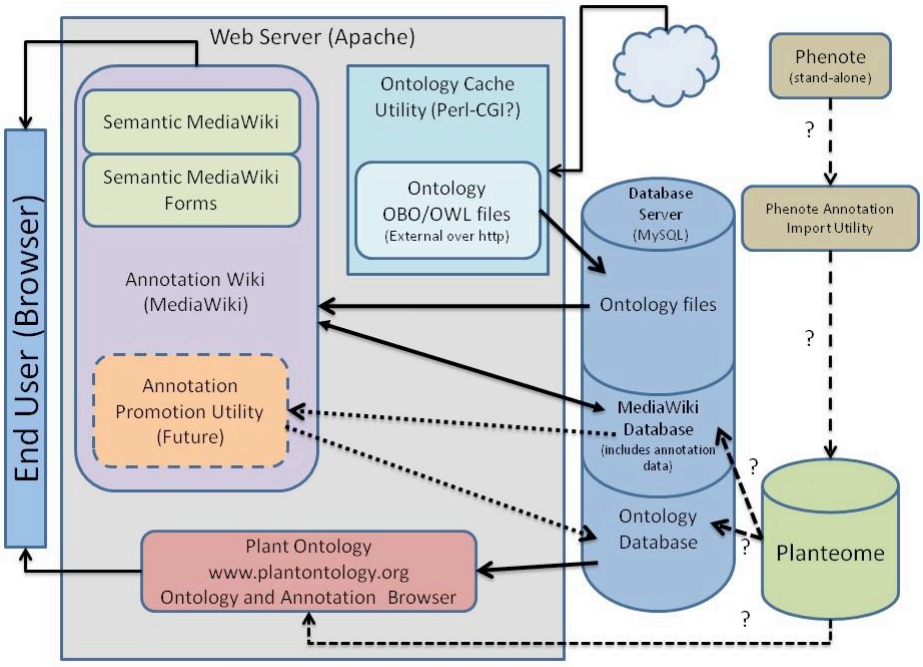

Figure 1: Architectural diagram of a semantic wiki site integrated with other systems (some existing, some proposed)

Methods

Features we have decided to explore in the early phase of this project include secure user accounts to track curator's individual annotations, a defined semantic structure for annotation data using RDF-XML[1] (*see Figure 2*), import and export utilities for tab-delimited data files (including GAF2

"association" files[2]) and relational databases, and external data sources for lists of species, genes, journals, *et al*. There will also be a focus on semantically query-able pages for reasoners and external semantic applications, cross-references to ontological data (including, but not limited to, the Plant Ontology[3], GO[4], and ChEBI[5]) and pre-loaded annotation data in the form of contributed association files (*e.g.* courtesy of TAIR[6], Gramene[7], MaizeGDB[8]). Finally, we would like to implement a curatorial approval process to move annotations from "proposed" to "approved" status; this is intended to create the appropriate balance between community and authority.

```
<!-- exported page data -->
<swivt:Subject rdf:about="http://palea.cgrb.oregonstate.edu/AnnoWiki/index.php/Special:URIResolver/O._sativa_-2D_Inflorescence">
  <rdfs:label>O. sativa - Inflorescence</rdfs:label>
  <swivt:page rdf:resource="http://palea.cgrb.oregonstate.edu/AnnoWiki/index.php/O._sativa_-_Inflorescence"/>
  <rdfs:isDefinedBy rdf:resource="http://palea.cgrb.oregonstate.edu/AnnoWiki/index.php/Special:ExportRDF/O._sativa_-_Inflorescence"/>
  <rdf:type rdf:resource="http://palea.cgrb.oregonstate.edu/AnnoWiki/index.php/Special:URIResolver/Category-3AAnnotations"/>
  <swivt:wikiNamespace rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">0</swivt:wikiNamespace>
  <property:Gene_ID rdf:datatype="http://www.w3.org/2001/XMLSchema#string">AWN 4</property:Gene_ID>
  <swivt:wikiPageModificationDate rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2010-10-14T21:38:53</swivt:wikiP
  <property:PO_Term rdf:datatype="http://www.w3.org/2001/XMLSchema#string">PO:0009049</property:PO_Term>
  <property:Species rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Oryza sativa</property:Species>
</swivt:Subject>
```

Figure 2: Sample RDF-XML data generated from prototype using the Semantic MediaWiki extension

Our platform stack consists of MediaWiki, PHP 5, MySQL 5.0 and the following extensions (at time of writing): Semantic MediaWiki[9] 1.5.2, Semantic Forms[10] 2.0.1, Data Transfer[11] 0.3.6, and External Data[12] 1.1 (*see Figure 1*).

Results (*see Figure 3*)

We intend to demonstrate that a semantic wiki providing convenient annotation entry, quality data, and publication/citation incentives will attract a broad and credible user base. It will be a valuable resource for researchers in botany, plant pathology, agriculture and -omics of all flavors. Use of a semantic data structure will encourage external querying of wiki annotation data and greater participation in the import and export of that data.

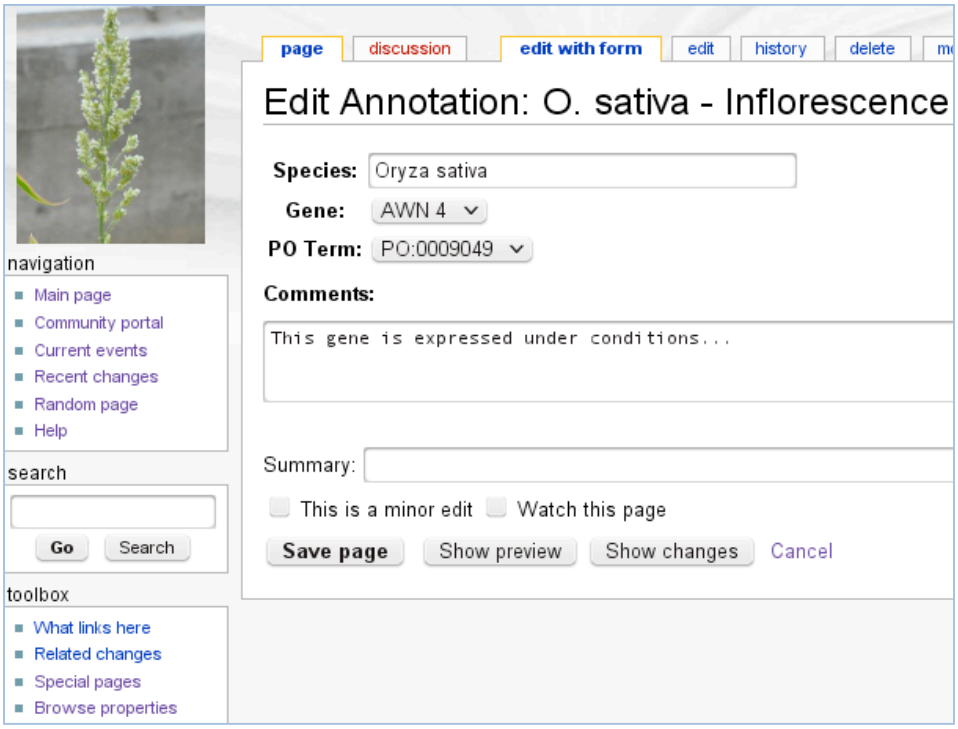

Figure 3: Prototype semantic wiki site (early user interface development using the Semantic Forms extension)

References

[1] RDF/XML Syntax Specification (Revised). W3C Recommendation 10 February 2004. http://www.w3.org/TR/REC-rdf-syntax/

[2] GO Annotation File Format 2.0 Guide. http://www.geneontology.org/GO.format.gaf-2_0.shtml

[3] Avraham S, Tung CW, Ilic K, Jaiswal P, Kellogg EA, McCouch S, Pujar A, Reiser L, Rhee SY, Sachs MM, Schaeffer M, Stein L, Stevens P, Vincent L, Zapata F, Ware D. The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. Nucleic Acids Res. 2008 Jan;36(Database issue):D449-54.

[4] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat. Genet.. May 2000;25(1):25-9.

[5] Degtyarenko, K., Hastings, J., de Matos, P., and Ennis, M. (2009). ChEBI: an open bioinformatics and cheminformatics resource. Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.], Chapter 14.

[6] Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, Radenbaugh A, Singh S, Swing V, Tissier C, Zhang P and Huala E. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Research 2008 36:D1009-D1014.

[7] Ware D, Jaiswal P, Ni J, Pan X, Chang K, Clark K, Teytelman L, Schmidt S, Zhao W, Cartinhour S, McCouch S, Stein L (2002) *Gramene: a resource for comparative grass genomics*. Nucleic Acids Research, 30, 103-105.

[8] Lawrence, CJ, Seigfried, TE, and Brendel, V. (2005) The Maize Genetics and Genomics Database. The community resource for access to diverse maize data. *Plant Physiology* 138:55-58.

[9] Semantic MediaWiki 1.5.2. http://semantic-mediawiki.org

[10] Semantic Forms Extension 2.0.1. http://www.mediawiki.org/wiki/Extension:Semantic_Forms

[11] Data Transfer Extension 0.3.6. http://www.mediawiki.org/wiki/Extension:Data_Transfer

[12] External Data Extension 1.2 http://www.mediawiki.org/wiki/Extension:External_Data