Semantic Annotation, Ontology Building, and Interactive Key Generation from Morphological Descriptions

Hong Cui,	University of Arizona
Alex Dusenbery	UMass-Boston
James Macklin	Agriculture Agri-Food, Canada
Fengqiong Huang	University of Arizona
Robert (Bob) Morris	UMass-Boston
Heather Cole	Agriculture Agri-Food, Canada

TDWG Beijing, Oct 22- 26 2012

Agenda

- Fine-Grained Semantic Markup
  - CharaParser, a semantic parser for phenotypes
    - Format conversion utilities: XML to RDF and SDD
  - OTO, a web-based application for ontology building
  - CFG, configurable field guide
- Challenges

CharaParser for Fine-Grained Semantic Annotation

- Annotate **factual** information from **textual morphological descriptions** of biodiversity in such a **detailed** manner that the machine readable annotation conveys the information of the original text.

- **Input**: plain text descriptions
- **Ouput**: XML files with organs/parts, characters, character states, relations, modifiers, and constraints explicitly annotated.
- Employs unsupervised machine learning and syntactic parsing techniques.

Example Annotations

human readable format

"Roots yellow to medium brown or black, thin."

- root color range= "yellow to medium brown"
- root color ="black"
- root extent="thin"

"Stems erect to prostrate, often with swollen nodes"

- stem orientation range= "erect to prostrate"
- node solid_shape = "swollen"
- Stems has_part nodes

Machine-Readable Format

- Native format: XML (eXtensible Markup Language)
  - XML schema for annotation (Web link)
- Example annotation in XML format
- XML annotation can be transformed into other standard formats, such as
  - SDD (Structured Descriptive Data)
  - RDF (Resource Description Framework)
  - Taxon-character matrices

CharaParser Performance

- Performance evaluated on
  - Flora of North America (FNA) volume 19
  - Treatise on Invertebrate Paleontology (TIP) part H.
  - Using precision (P) and recall (R)
- Results published in JASIST 2012.

	N (complex)	Structure P/R	Character P/R	Relation P/R	Sentence P/R	Correct N
FNA.v19	559 (22%)	99/95	91/90	85/49	96/94	434(78%)
TIP.h	457(42%)	97/97	80/87	79/59	90/90	262 (57%)

From Semantic Annotation to Taxon-Character Matrices

- The XML schema for annotation is based on entity-relation model
  - Entity, attributes, and relations among entities
- Translating XML annotation to taxon-character matrices is straightforward, but
  - Sparse matrix problem
    - Inherit characters/character states from higher taxa
  - Polymorphism
  - Numerical expressions
  - Parenthetical characters

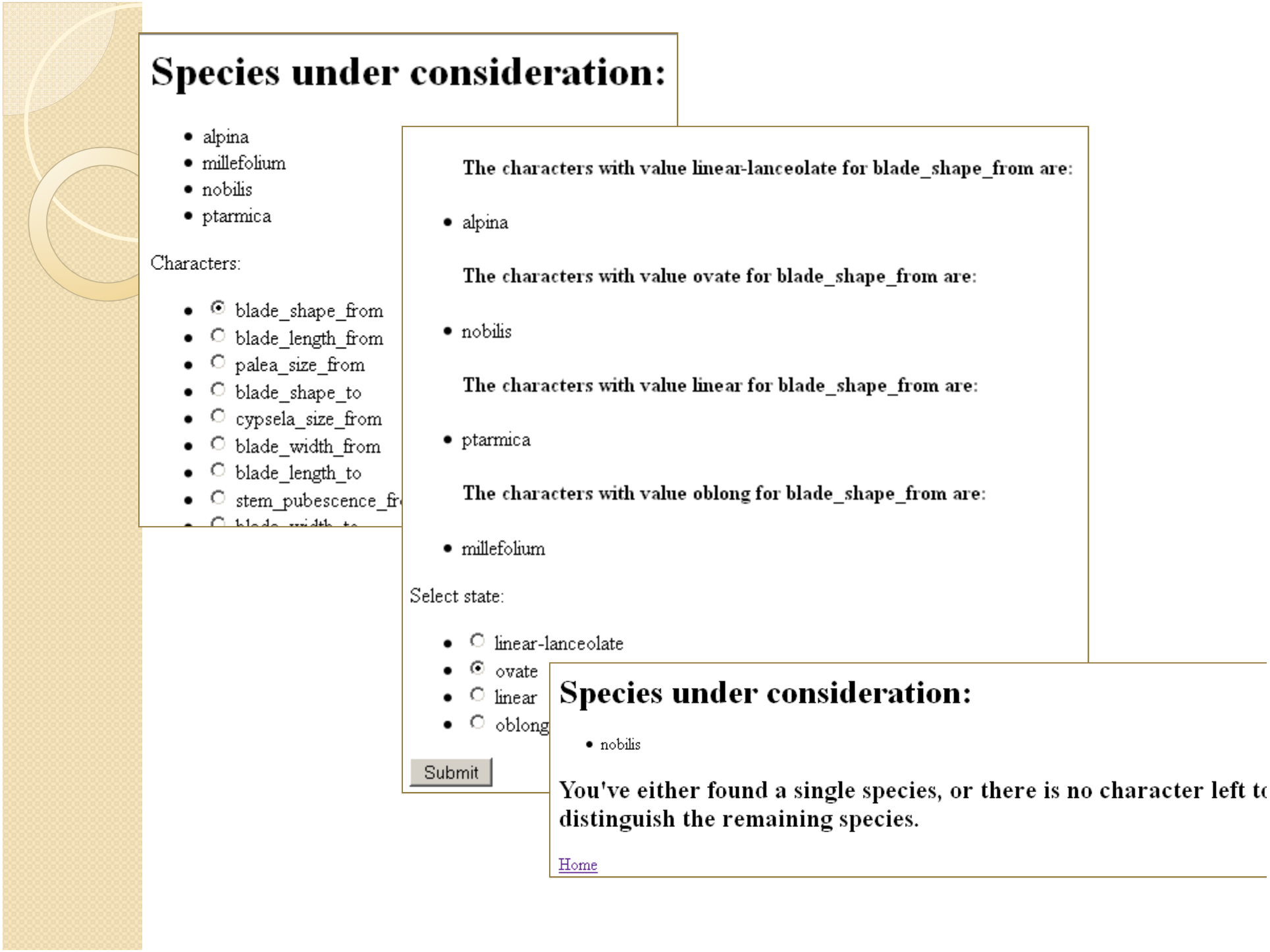
Matrix Generated for ACHILLEA (Plant Genus)

inheritance | polymorphism | range value | range value

	apex_texture	array_architecture	blade_length_from	blade_length_to	blade_shape_from	blade_shape_to
ACHILLEA	scarious	compact|open|simple|compound|corymbiform			linear|oblong-lanceolate|lobed	oblong-lanceolate|linear|lobed
alpina	scarious	crowded|simple|compound|corymbiform	0.05	0.1	linear-lanceolate|oblong-lanceolate	oblong-lanceolate|linear-lanceolate
millefoliu m	scarious	simple|compound|corymbiform	0.035	0.35	oblong|lanceolate|lobed	oblong|lanceolate|lobed
nobilis	scarious	simple|compound|corymbiform	0.015	0.03	ovate|lobed	ovate|lobed
ptarmica	scarious	simple|compound|corymbiform	0.003	0.01	linear|lanceolate	narrowly lanceolate|linear|lanceolate

Configurable Field Guide

- Created by Robert Morris and students.
- **Input**: taxon-character matrix
- **Output**: interactive field guide
- Ranks characters in their ability to evenly partition the taxa in question.
  - At each partition, maximize information gain.
- Will integrate a new information entropy-based algorithm that deals with polymorphism (Wang, Z, 2012).


Species under consideration:
alpina
millefolium
nobilis
ptarmica
Characters:
blade_shape_from
blade_length_from
palea_size_from
blade_shape_to
cypsela_size_from
blade_width_from
blade_length_to
The characters with value linear-lanceolate for blade_shape_from are:
alpina
The characters with value ovate for blade_shape_from are:
nobilis
The characters with value linear for blade_shape_from are:
ptarmica
The characters with value oblong for blade_shape_from are:
millefolium
Select state:
linear-lanceolate
ovate
linear
Submit
Species under consideration:
nobilis
You've either found a single species, or there is no character left t
distinguish the remaining species.
Home

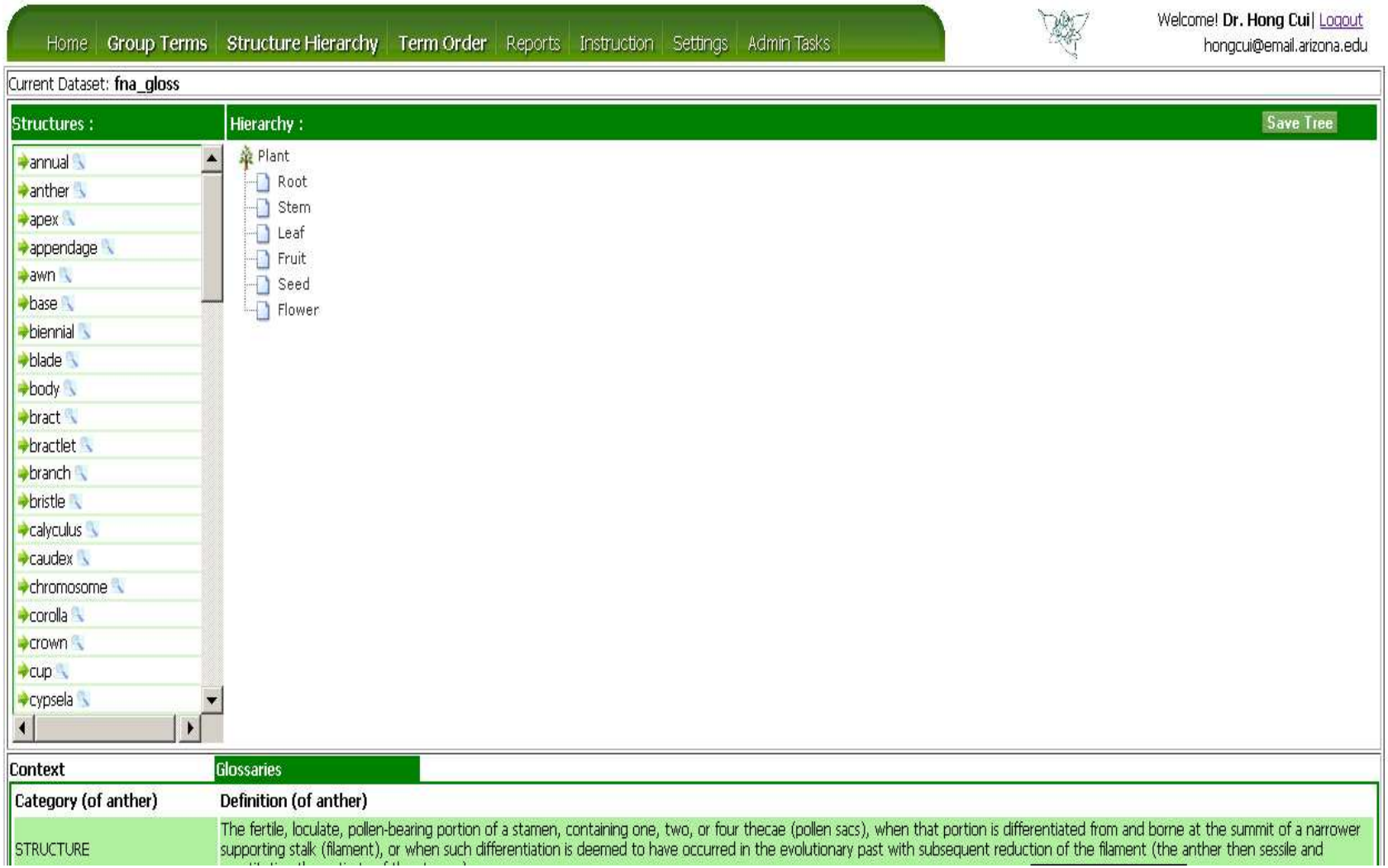
OTO: Ontology Term Organizer

- http://biosemantics.arizona.edu/ONTNEW/
  - username : OTOdemo
  - password: OTOdemopass
  - Select demo dataset: OTO_demo.

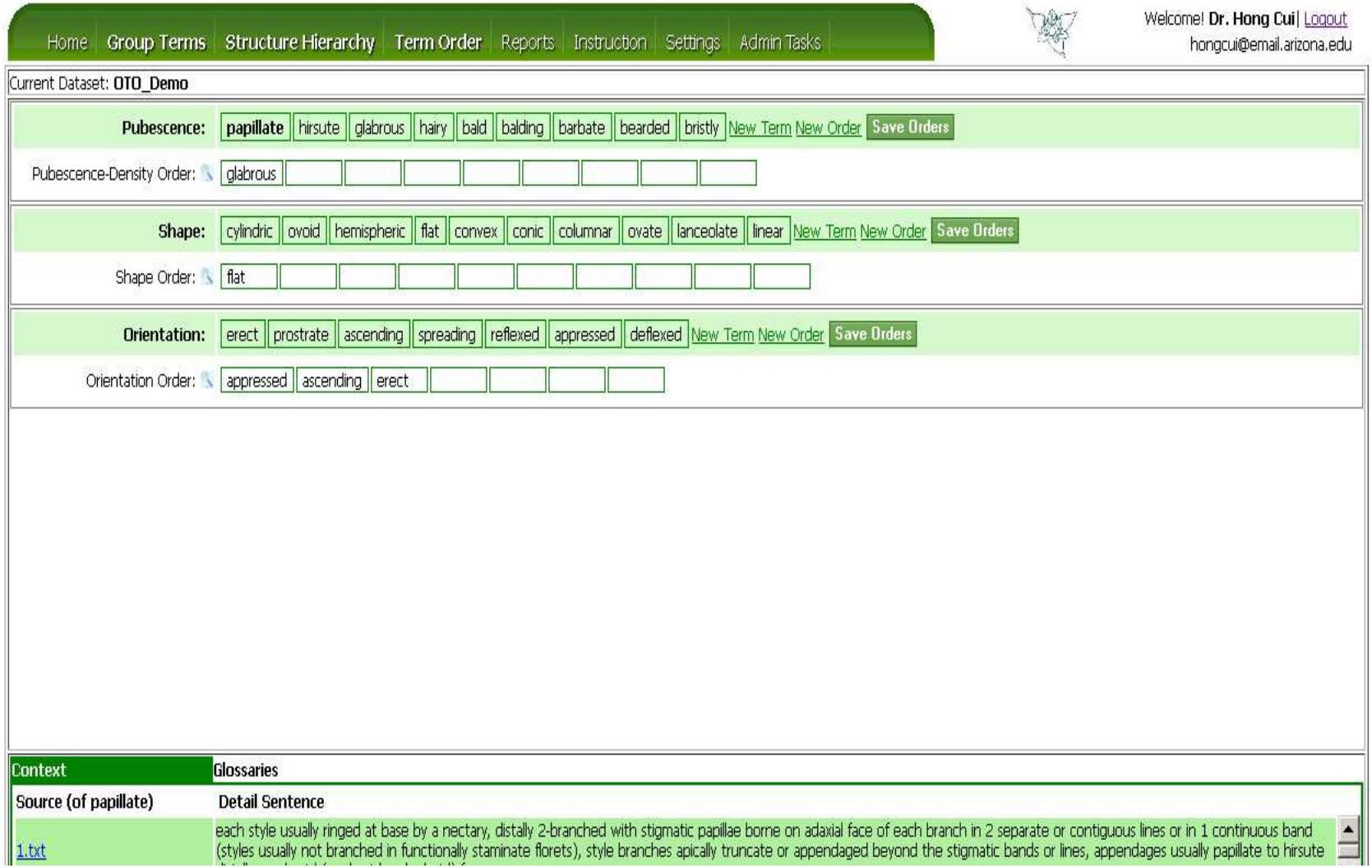
- Web-based application
  - **Input**: terms/expressions (extracted by CharaParser)
  - **Output**: clean glossaries or raw ontologies.
  - Used by biologists
  - Sort is_a, part_of and order relationships of terms extracted by CharaParser
  - Resolve conflicts based on consensus

OTO: Group Terms

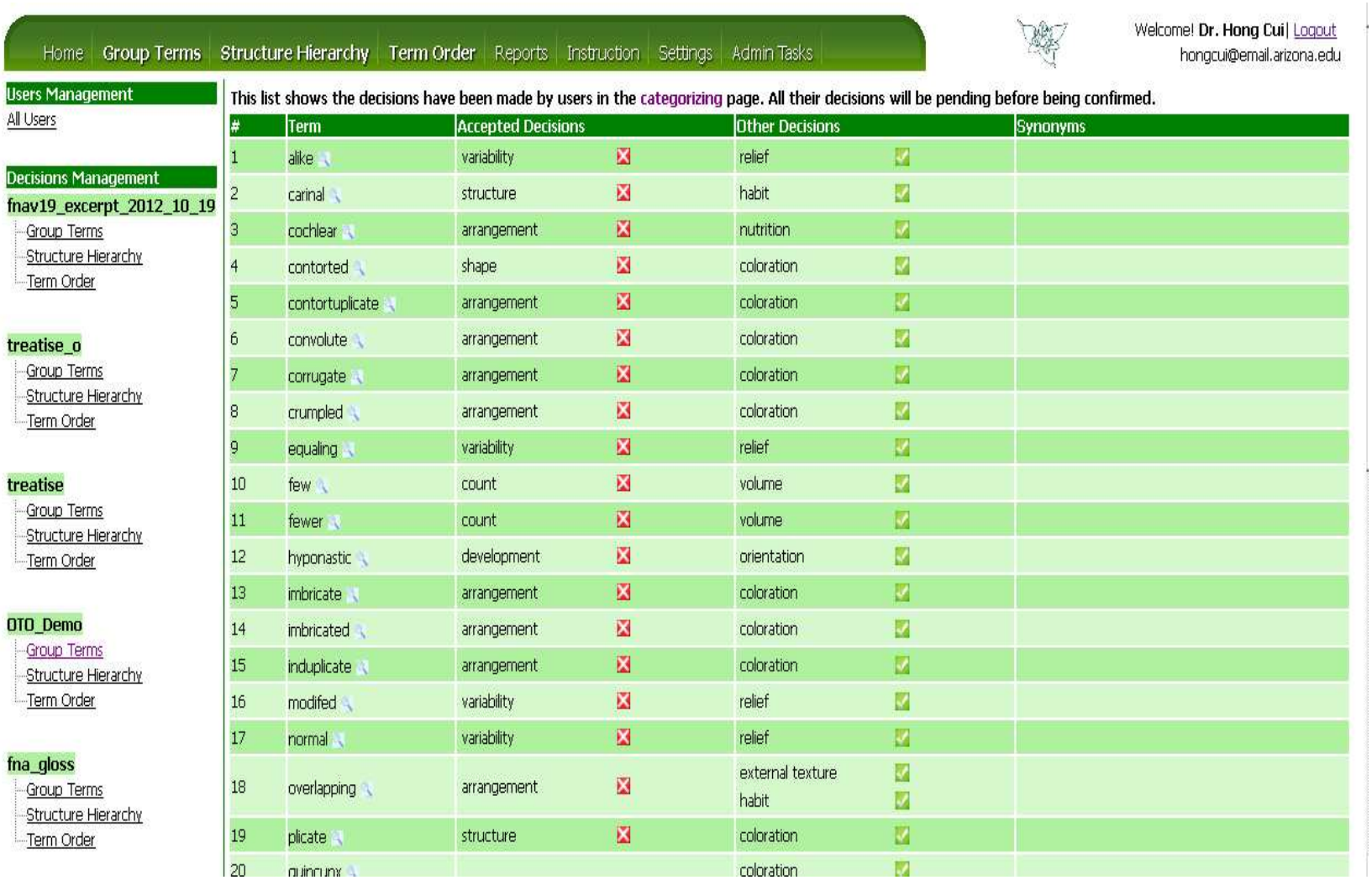
OTO: Structure Hierarchy



OTO: Term Order

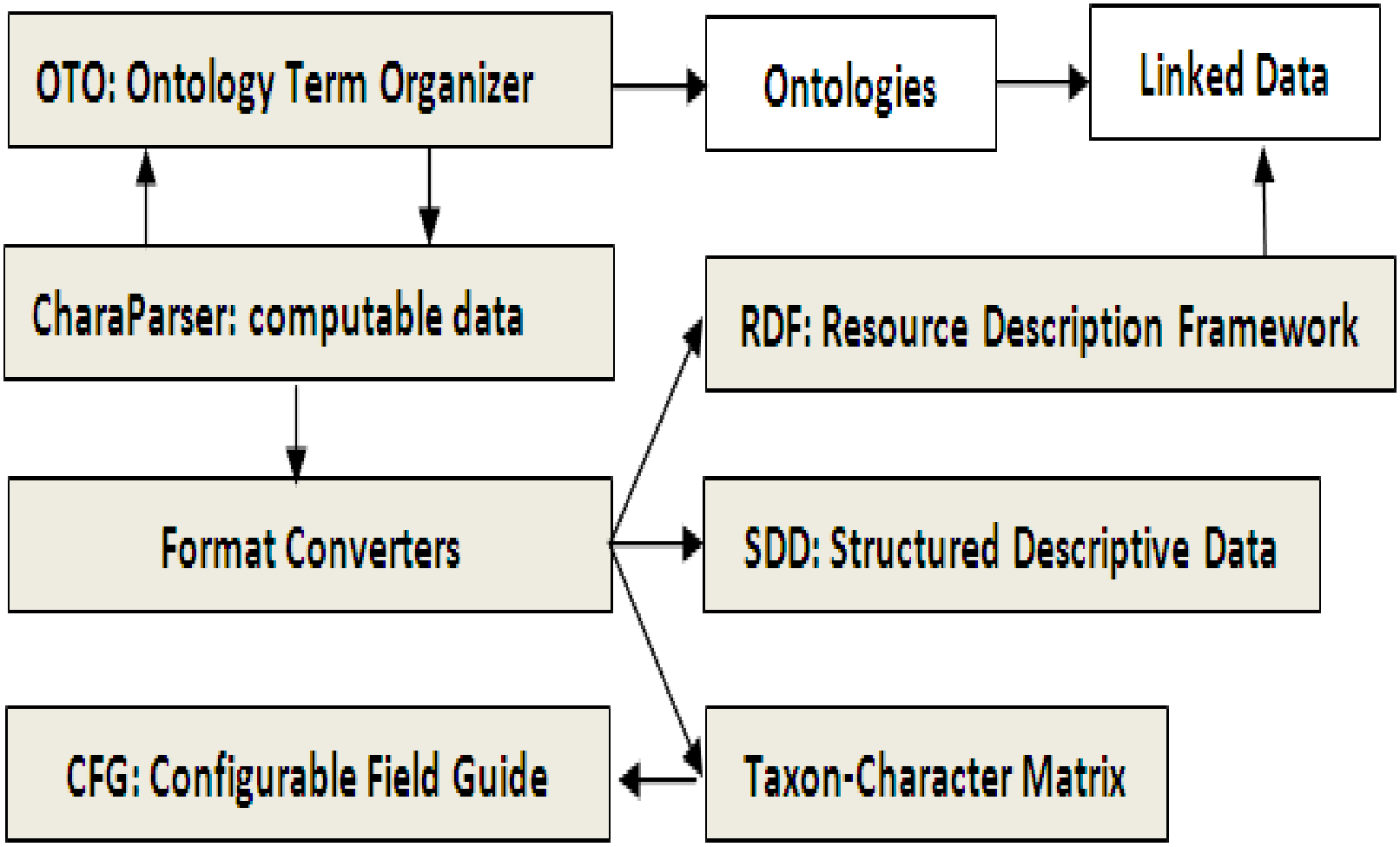


OTO: Admin Tools



#	Term	Accepted Decisions	Other Decisions	Synonyms
1	alike	variability	relief	
2	carinal	structure	habit	
3	cochlear	arrangement	nutrition	
4	contorted	shape	coloration	
5	contortuplicate	arrangement	coloration	
6	convolute	arrangement	coloration	
7	corrugate	arrangement	coloration	
8	crumpled	arrangement	coloration	
9	equaling	variability	relief	
10	few	count	volume	
11	fewer	count	volume	
12	hyponastic	development	orientation	
13	imbricate	arrangement	coloration	
14	imbricated	arrangement	coloration	
15	induplicate	arrangement	coloration	
16	modifed	variability	relief	
17	normal	variability	relief	
18	overlapping	arrangement	external texture habit	
19	plicate	structure	coloration	
20	quincunx		coloration	

Summary


OTO: Ontology Term Organizer
Ontologies
Linked Data
CharaParser: computable data
RDF: Resource Description Framework
Format Converters
SDD: Structured Descriptive Data
CFG: Configurable Field Guide
Taxon-Character Matrix

Challenges

- Adapting CharaParser from semi-structured sublanguage to natural language and mixed style input
- Building extension ontologies for target domains
- Determining the appropriate level of expressiveness in annotation: how fine-grained is fine enough
- Measuring character-based semantic similarity among taxa

Acknowledgements

- NSF award no. EF-0849982
- Flora of North America Project

- Contact info:
  - hongcui@email.arizona.edu

More Annotation Examples

1. Example, Num-Example (fna, complex)
2. Constraint-Example (Treatise.h, complex)
3. Num-Example (BHL-OCR, simple)
4. Example (Ant-OCR, complex)