



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
- **Output:** 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 linear-lanceolate
millefolium	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
- stem_pubescence_fr
- blade_width_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
- oblong

Submit

Species under consideration:

- nobilis

You've either found a single species, or there is no character left to distinguish the remaining species.

[Home](#)



OTO: Ontology Term Organizer

- <http://biosemantics.arizona.edu/ONTNEW/>
 - username : OTOfdemo
 - password: OTOfdemopass
 - Select demo dataset: OTOf_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

Home Group Terms Structure Hierarchy Term Order Reports Instruction Settings Admin Tasks

Welcome! [Dr. Hong Cui](#) | [Logout](#)
hongcui@email.arizona.edu

Current Dataset: **fna_gloss** (2738 terms, 2542 reviewed)

Terms: [Save Decisions/Submit Review History](#) [New Category](#)

Terms:	Categories:					
<input type="checkbox"/> netted_3 <input type="checkbox"/> compact_1 <input type="checkbox"/> congested_1 <input type="checkbox"/> lax_1 <input type="checkbox"/> sparser_1 <input type="checkbox"/> rust_1 <input type="checkbox"/> contorted_2 <input type="checkbox"/> acerose_1 <input type="checkbox"/> contractile_1 <input type="checkbox"/> pleurorhizal_1 <input type="checkbox"/> notorhizal_1 <input type="checkbox"/> diplocolobal_1 <input type="checkbox"/> false_1 <input type="checkbox"/> funicular_1 <input type="checkbox"/> integumentary_1 <input type="checkbox"/> adventitious_1 <input type="checkbox"/> involucre_1 <input type="checkbox"/> raphal_1	arrangement accumbent acyclic adjacent aggregated aligned alternate alternating apart approximate	coloration aciculate ashy banded beige bicolor bicolorous black blackening blackish	condition desiccated fresh healthy intact sclerified undamaged worn disintegrating_1 withered_1	count abundant copious decreasing few fewer many multiple none numerous	course arcuate curling flexuose flexuous S-shaped sigmoid sinuous spiraled spiraling	dehiscence circumscissile dehiscent dehiscent disintegrating explosive extrorse indehiscence indehiscent introrse
	density	depth	derivation	development	duration	external texture
	exudation	fixation	fragility	fusion	germination	habit
	height	internal texture	length	life_stage	life_style	location
	maturation	nutrition	odor	orientation	origin	pattern
	position	prominence	reflectance	relief	reproduction	shape
	size	structure	taste	variability	venation	vernation
	volume	width				

Locations Context Glossaries

Source (of aggregated)	Detail Sentence
2.txt	borne singly (sometimes on scapiform stems) or in corymbiform, paniculiform, or racemiform arrays (aggregated in second-order heads, florets 1-3 per individual head in Hecastocleis).
228.txt	Heads usually heterogamous (usually radiate) [homogamous, discoid], borne singly (on scapiform peduncles) [in corymbiform, racemiform, or umbelliform arrays, sometimes aggregated in second-order heads].

OTO: Structure Hierarchy

Home Group Terms Structure Hierarchy Term Order Reports Instruction Settings Admin Tasks

Welcome! **Dr. Hong Cui** [Logout](#)
hongcui@email.arizona.edu

Current Dataset: **fna_gloss**

Structures : **Hierarchy :** [Save Tree](#)

- annual
- anther
- apex
- appendage
- awn
- base
- biennial
- blade
- body
- bract
- bractlet
- branch
- bristle
- calyculus
- caudex
- chromosome
- corolla
- crown
- cup
- cypsela

Plant

- Root
- Stem
- Leaf
- Fruit
- Seed
- Flower

Context	Glossaries
Category (of anther)	Definition (of anther)
STRUCTURE	The fertile, loculate, pollen-bearing portion of a stamen, containing one, two, or four thecae (pollen sacs), when that portion is differentiated from and borne at the summit of a narrower supporting stalk (filament), or when such differentiation is deemed to have occurred in the evolutionary past with subsequent reduction of the filament (the anther then sessile and

OTO: Term Order

[Home](#) [Group Terms](#) [Structure Hierarchy](#) [Term Order](#) [Reports](#) [Instruction](#) [Settings](#) [Admin Tasks](#)



Welcome! [Dr. Hong Cui](#) | [Logout](#)
hongcui@email.arizona.edu

Current Dataset: **OTO_Demo**

Pubescence: [papillate](#) [hirsute](#) [glabrous](#) [hairy](#) [bald](#) [balding](#) [barbate](#) [bearded](#) [bristly](#) [New Term](#) [New Order](#) [Save Orders](#)

Pubescence-Density Order: [glabrous](#)

Shape: [cylindric](#) [ovoid](#) [hemispheric](#) [flat](#) [convex](#) [conic](#) [columnar](#) [ovate](#) [lanceolate](#) [linear](#) [New Term](#) [New Order](#) [Save Orders](#)

Shape Order: [flat](#)

Orientation: [erect](#) [prostrate](#) [ascending](#) [spreading](#) [reflexed](#) [appressed](#) [deflexed](#) [New Term](#) [New Order](#) [Save Orders](#)

Orientation Order: [appressed](#) [ascending](#) [erect](#)

Context [Glossaries](#)

Source (of papillate)

[1.txt](#)

Detail Sentence

each style usually ringed at base by a nectary, distally 2-branched with stigmatic papillae borne on adaxial face of each branch in 2 separate or contiguous lines or in 1 continuous band (styles usually not branched in functionally staminate florets), style branches apically truncate or appendaged beyond the stigmatic bands or lines, appendages usually papillate to hirsute

OTO: Admin Tools

[Home](#)
[Group Terms](#)
[Structure Hierarchy](#)
[Term Order](#)
[Reports](#)
[Instruction](#)
[Settings](#)
[Admin Tasks](#)



Welcome! [Dr. Hong Cui](#) | [Logout](#)
hongcui@email.arizona.edu

Users Management

[All Users](#)

Decisions Management

[fnav19_excerpt_2012_10_19](#)

[Group Terms](#)
[Structure Hierarchy](#)
[Term Order](#)

[treatise_o](#)

[Group Terms](#)
[Structure Hierarchy](#)
[Term Order](#)

[treatise](#)

[Group Terms](#)
[Structure Hierarchy](#)
[Term Order](#)

[OTO_Demo](#)

[Group Terms](#)
[Structure Hierarchy](#)
[Term Order](#)

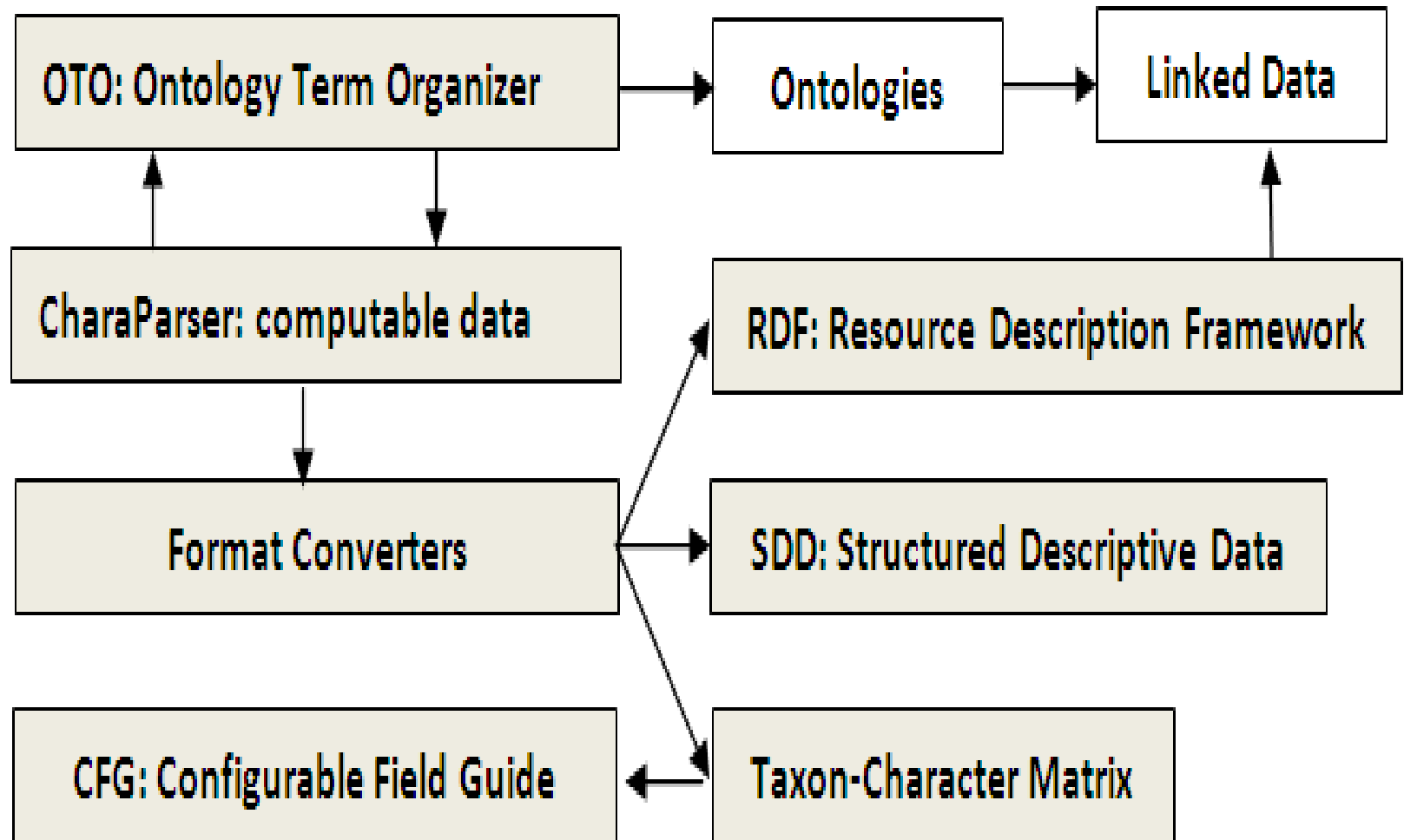
[fna_gloss](#)

[Group Terms](#)
[Structure Hierarchy](#)
[Term Order](#)

This list shows the decisions have been made by users in the [categorizing](#) page. All their decisions will be pending before being confirmed.

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

Summary





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)