SPNHC BRIDGING CONTINENTS 2009

The Apiary Project

High-Throughput Workflow for Computer-Assisted Human Parsing of Biological Specimen Label Data

Jason Best, Amanda Neill – BRIT

Dr. Bill Moen - UNT

Funded by Institute of Museum and Library Services National Leadership Grant # 06-08-0079-08

The Apiary Project is a collaboration of the Botanical Research Institute of Texas and the University of North Texas.

BRIT – Jason Best, Amanda Neill, Dr. Brooke Byerley, Becky Brandenburg, Tiana Franklin, Keri McNew, Sean Murphy

UNT - Dr. Bill Moen, Dr. Miguel Ruiz, Ali Amin, Zainab Arsiwala, Jane Huang, Melody McCotter

The Challenge

- Millions of specimens with valuable data that are inaccessible in analog form
- Average of 5 minutes for one person to type the label data into their respective fields for one specimen. This would require about 40 years for 1 million specimens.

Collector Name	Number	Scientific Name
J. C Taylor	31007	Solidago altiplanities
W. Hess	7283	Spiraea densiflora
H. St. John	14846	Sophora manejarevaensis
C.G. Pringle	13949	Polygala minutifolia
R.F. Hoover	3614	Senecio clevelandii
O. Degener	33,680	Wikstroemia perdita

The Solution – Two Heads, Many Hands

Computers and humans working together

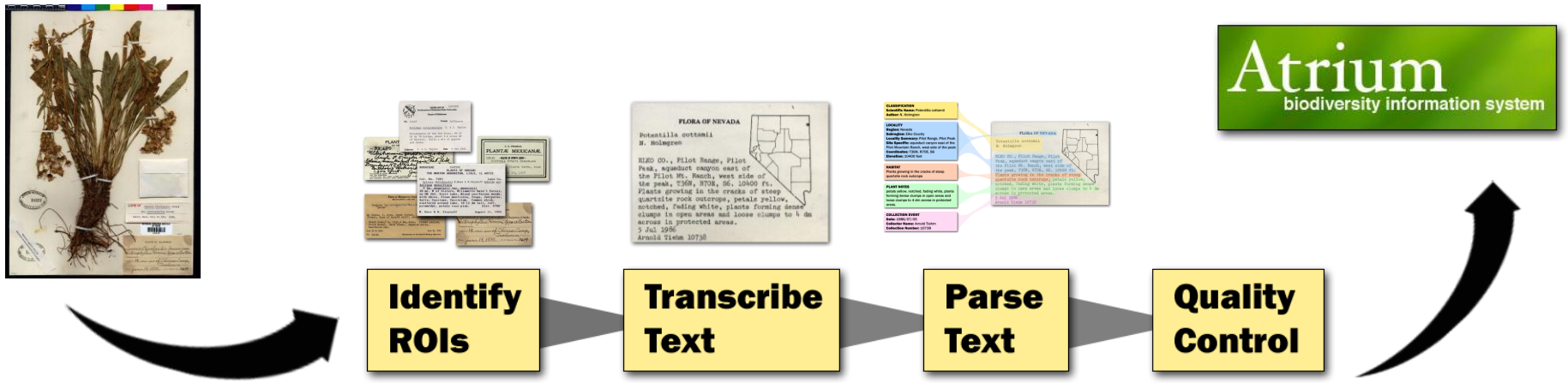
"Two heads are better than one."
– John Heywood

Many humans collaborating through a web-based system

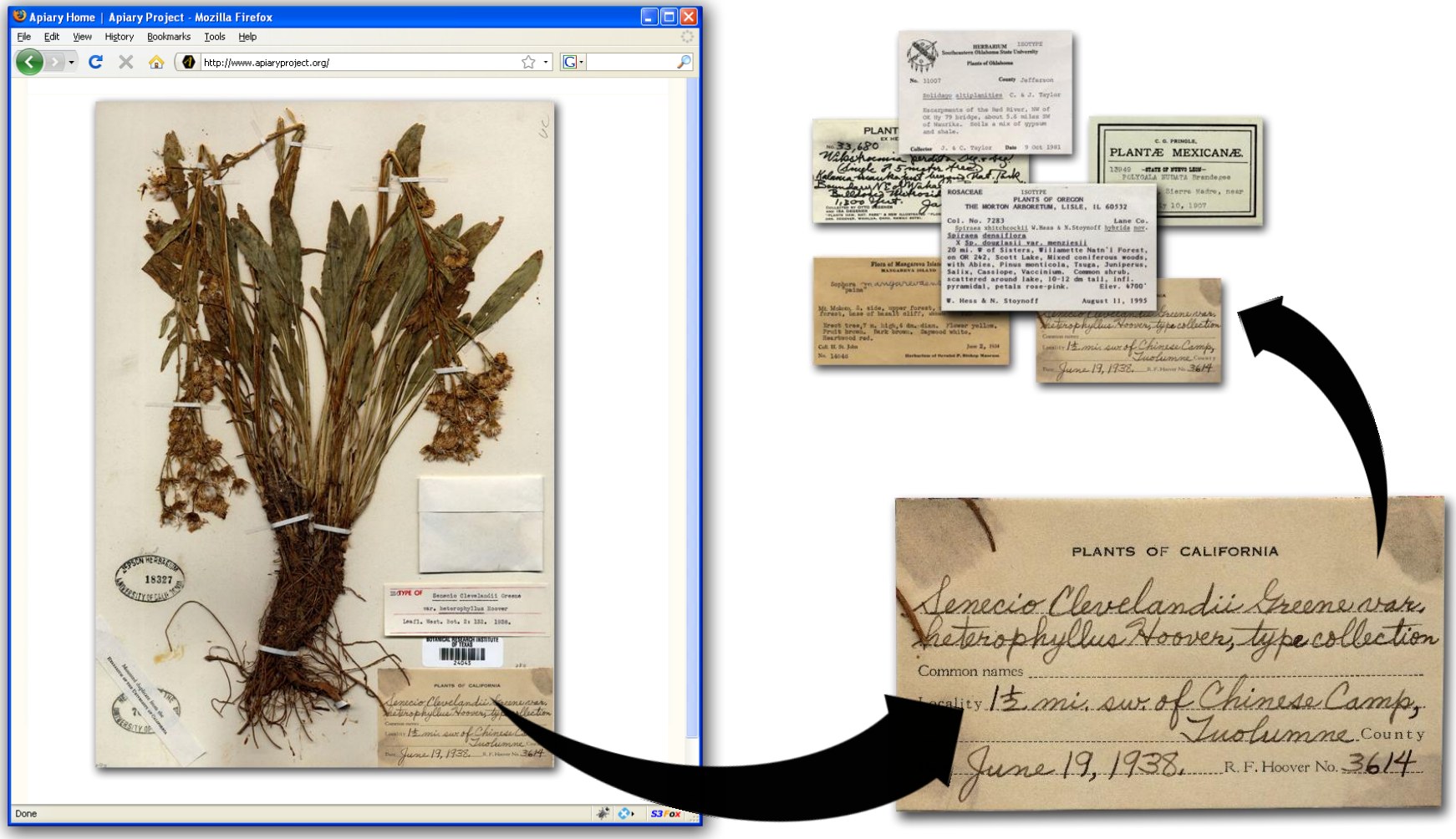
"Many hands make light work."
– John Heywood

The Technology and Workflow

- Speeds the process of extracting textual data from herbarium specimens to transform them into standardized data formats to facilitate sharing and improve access
- A web-based application accessible at any time, anywhere
- Leverages computer processes to extract text and parse data when possible
- Allows many individuals to participate in various stages of transforming specimen data through optimized interfaces



Identify Regions of Interest on specimen



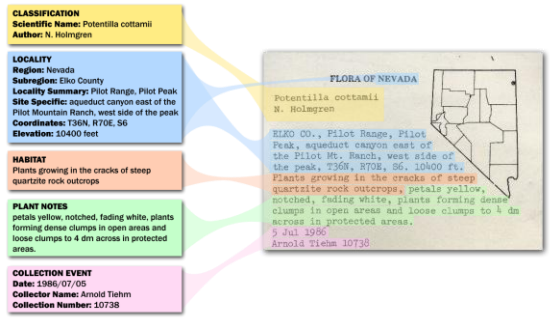
Transcribe and Parse Text

Region of interest images

Digital text transcribed by human or machine

FLORA OF NEVADA
Potentilla cottamii
N. Holmgren

ELKO CO., Pilot Range, Pilot
Peak, aqueduct canyon east of the Pilot Mt. Ranch, west side of the peak, BSN, R7012, S6. 10400 ft. Plants growing in the cracks of steep quartzite rock outcrops, petals yellow, notched, fading white, plants forming dense clumps in open areas and loose clumps to D dm across in protected areas.
5 Jul 1986
Arnold Tiehm 10738



Text parsed by human or machine

Collector Name	Number	Scientific Name
J. C Taylor	31007	Solidago altiplanities
W. Hess	7283	Spiraea densiflora
H. St. John	14846	Sophora manejarevaensis
C.G. Pringle	13949	Polygala minutifolia
R.F. Hoover	3614	Senecio clevelandii
O. Degener	33,680	Wikstroemia perdita

Quality control and submission to database