Reaching Consensus in Crowdsourced Transcription of Biocollections Information

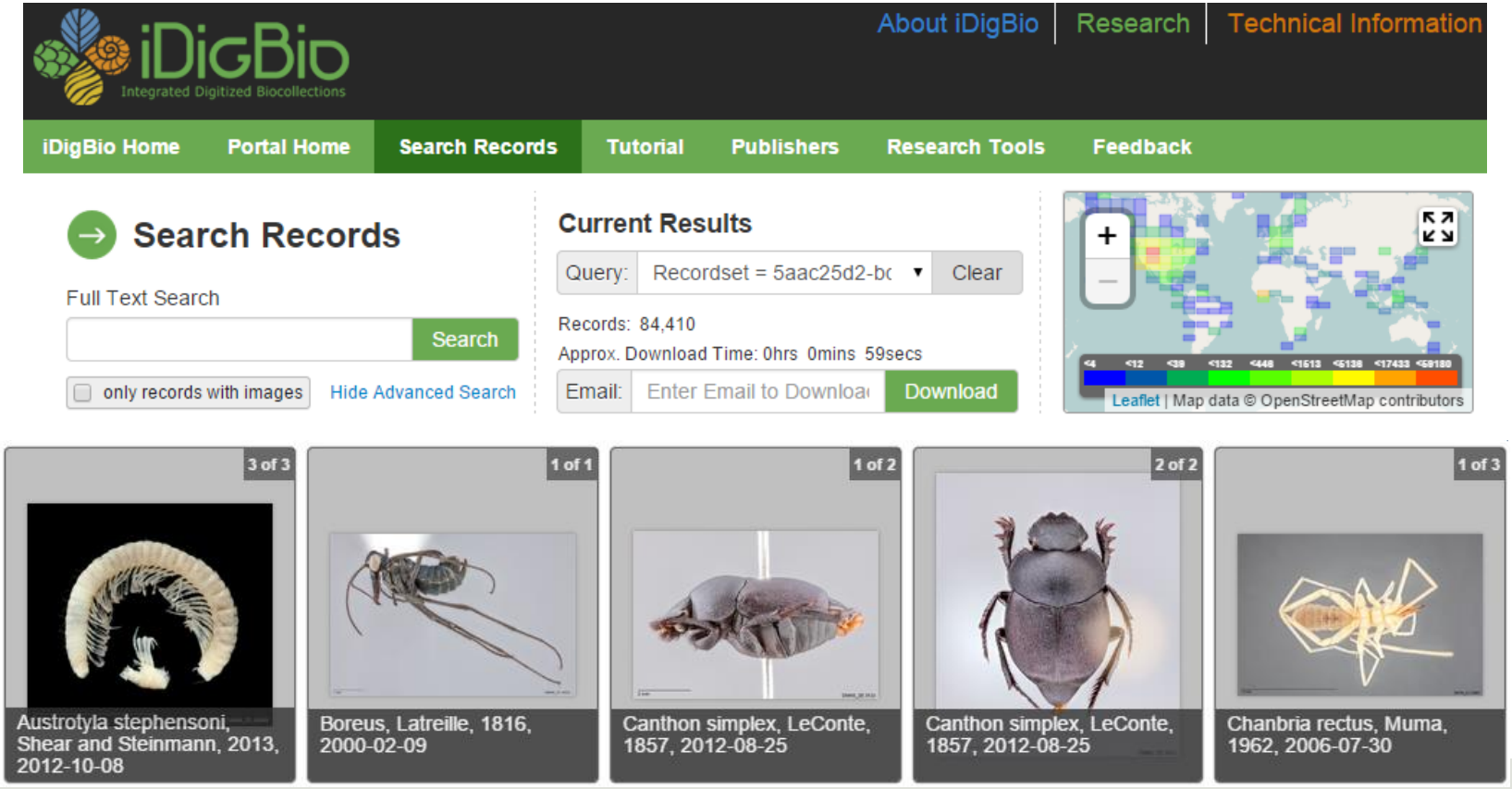
Andréa Matsunaga (ammatsun@ufl.edu), Austin Mast, and José A.B. Fortes

10th IEEE International Conference on e-Science
October 23, 2014
Guarujá, SP, Brazil

iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Background and Motivation

- Estimated 1-billion biological specimens in the US
- iDigBio + Thematic Collections Network (TCN) + Partners to Existing Network (PEN)



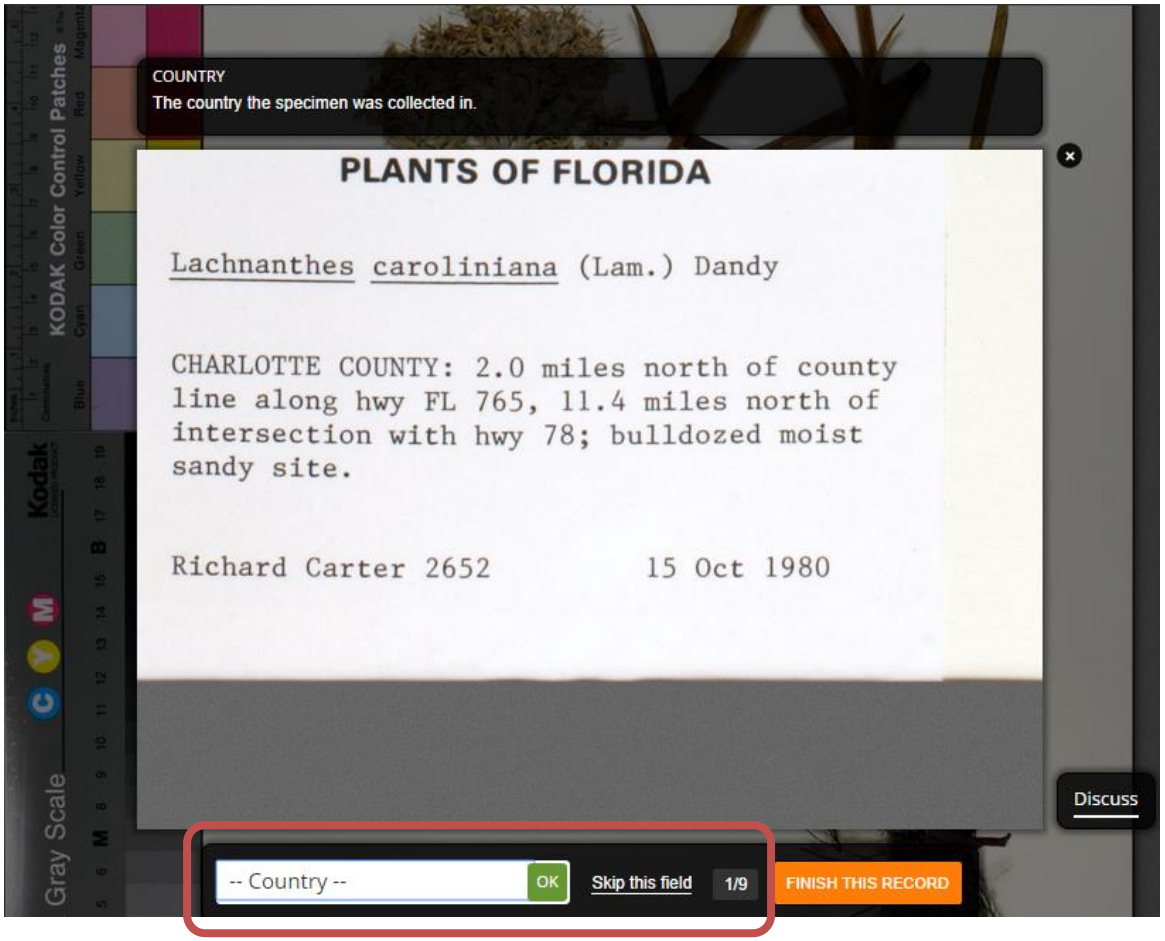
Background and Motivation

- TCNs and PENs performing digitization work
- Generating images, transcribing information about what/when/where/who

- If digitization took 1 second and if we performed in sequence:
  - 1,000,000,000 seconds > 30 years
- Parallelism:
  - Crowdsourcing!!

Crowdsourcing Transcription Projects

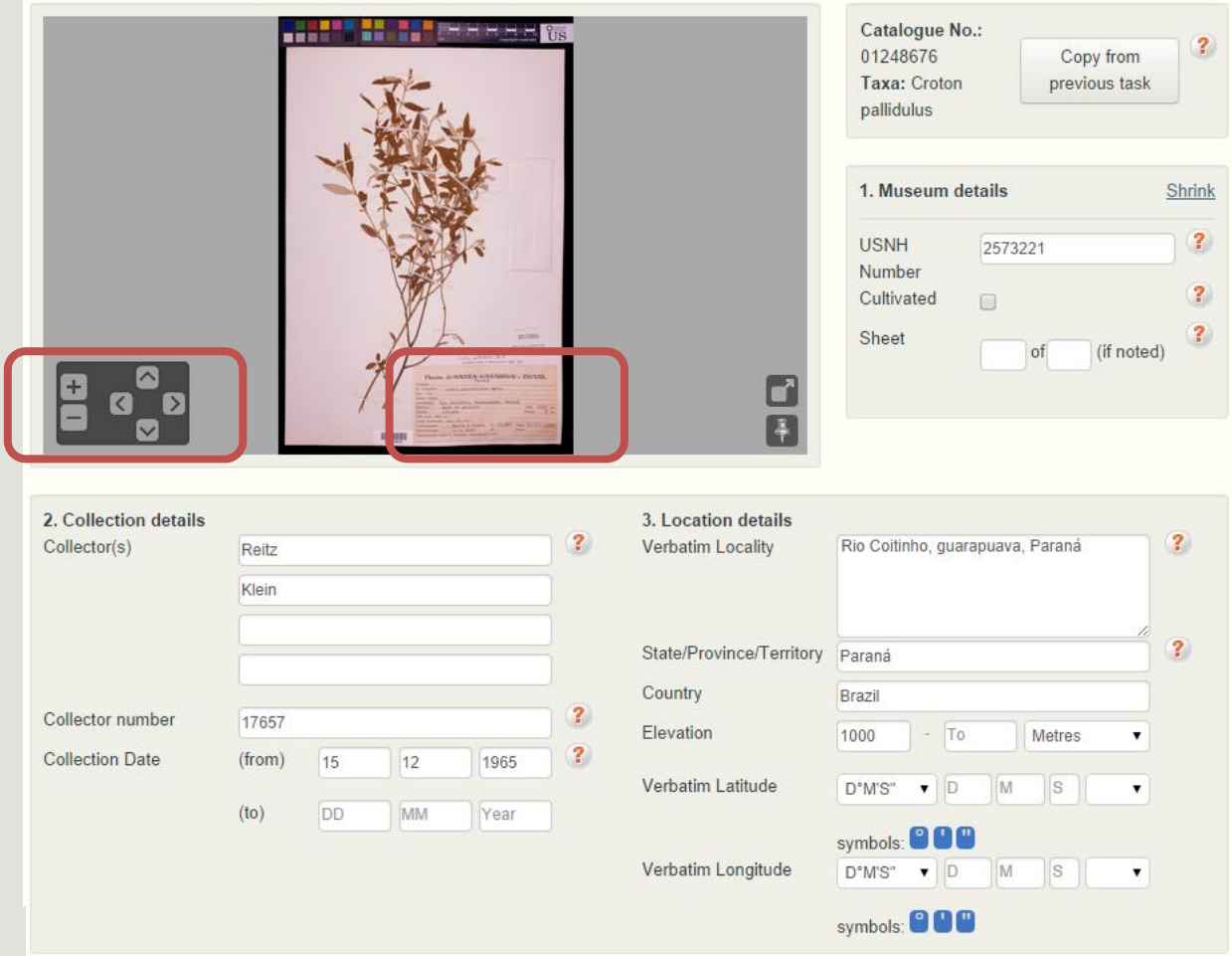
- NotesFromNature (http://www.notesfromnature.org/)
  - Zooniverse platform



- Once the user selects the region with the label, s/he can start transcribing and parsing information to a number of pre-defined fields
- For a requester, a pre-defined number of transcriptions are returned

Crowdsourcing Transcription Projects

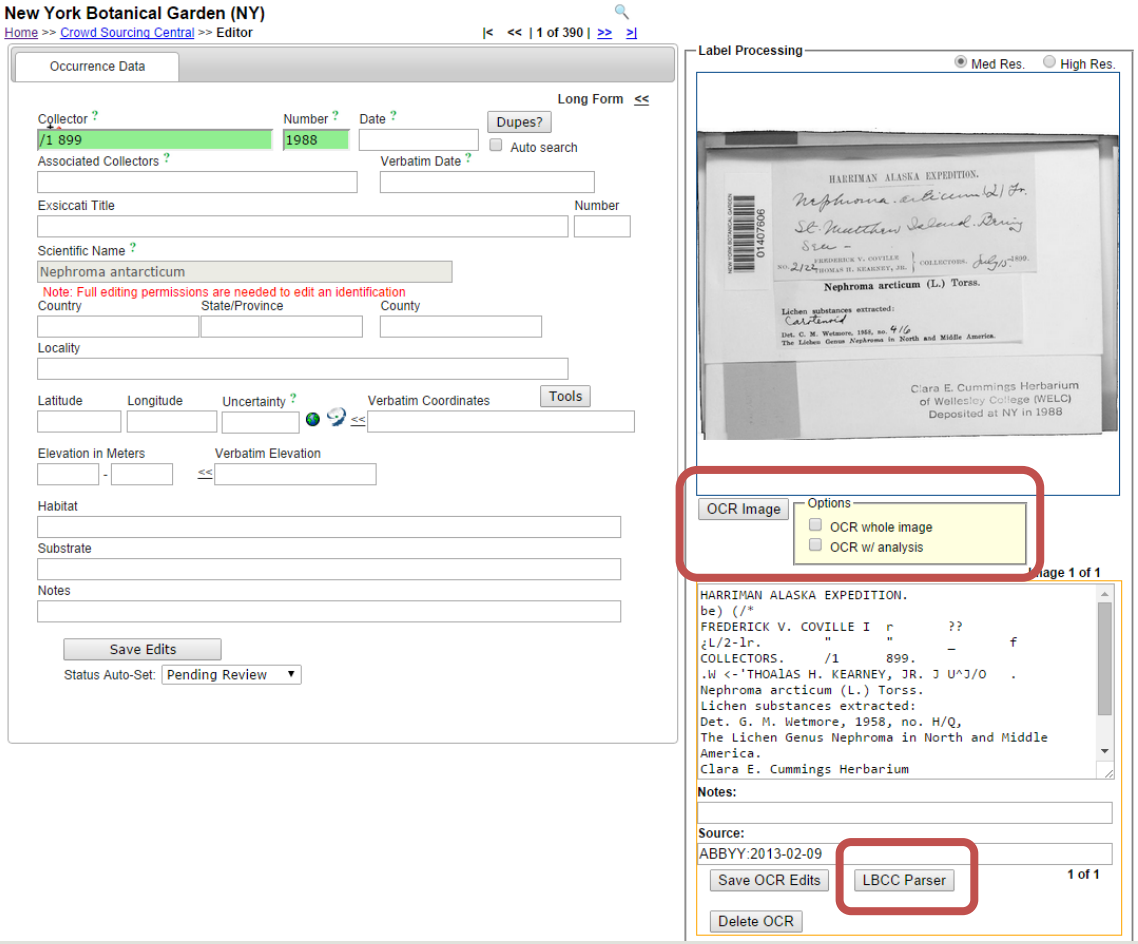
- ALA (http://volunteer.ala.org.au/)
  - Platform: Grails



- User zooms in to read the label and parse to the custom pre-defined terms
- Single worker followed by expert approval

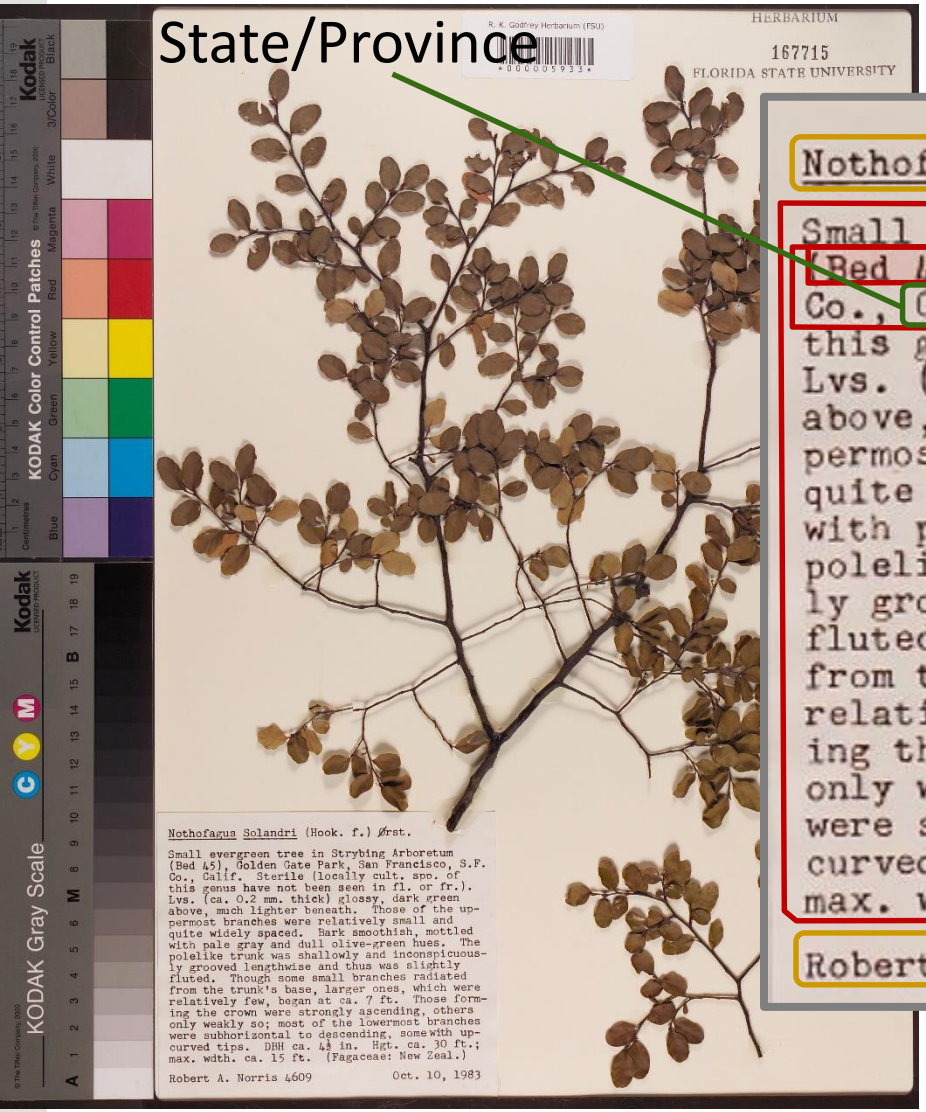
Crowdsourcing Transcription Projects

- Symbiota (http://lbcc1.acis.ufl.edu/volunteer)
  - Platform: PHP



- Ability to OCR and parse data
- Single worker followed by expert approval

The transcription task



Scientific Name

Scientific Author

Location

County

State/Province

Collector Number

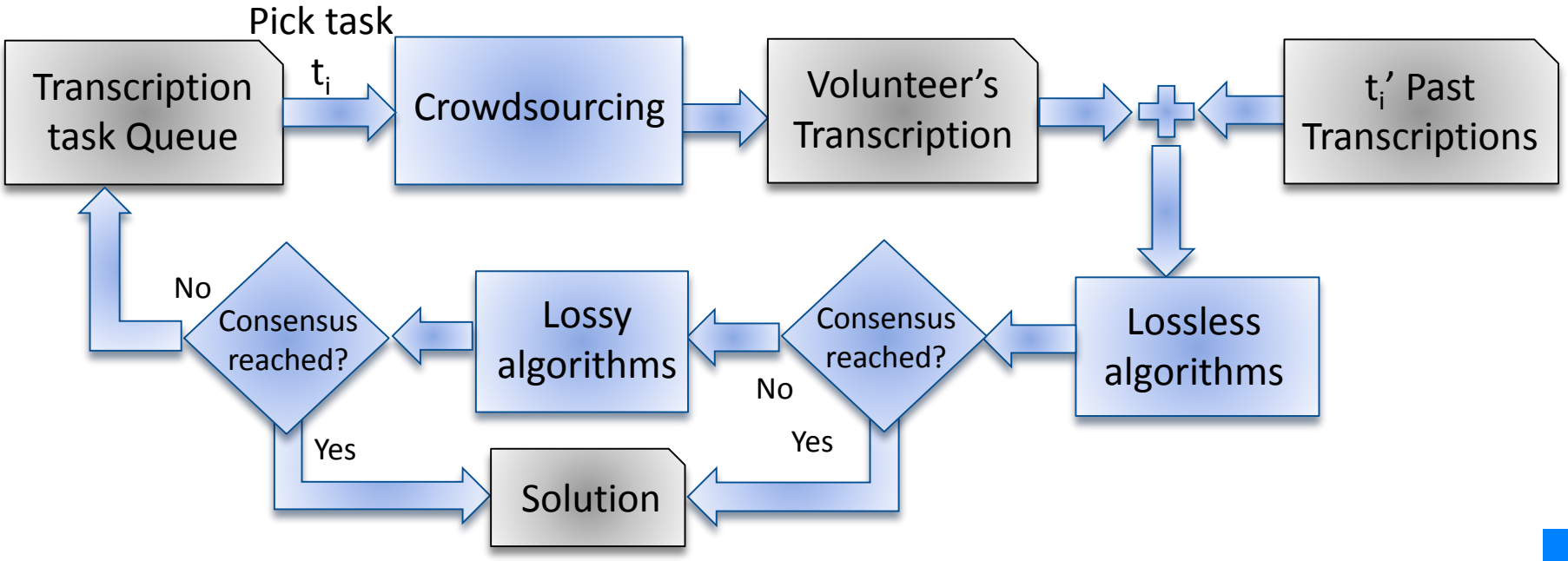
Collection Date

Collected by

Habitat and description

Proposed Consensus Approach

- Goal:
  - Reach consensus with minimum number of transcriptions
- Method:
  - Control the number of workers per task
  - Apply lossless and/or lossy algorithms per field



Lossless normalization algorithms

Code	Lossless functionality
b	Removes all extra whitespaces
n	Apply specific transformation functions on a per-field basis (e.g., to normalize section/township/range, proper names, and latitude/longitude)
t	Apply specific translation tables on a per-field basis to expand abbreviations (e.g., hy, hwy, and hiway to highway) or to shorten expansions (e.g., Florida to FL)

Lossy normalization algorithms

Code	Lossy functionality
w	Approximate comparison by ignoring all whitespace (e.g., “0–3” is equivalent to “0 – 3”)
c	Case insensitive approximate comparison (e.g., “Road” and “road” are considered equivalent)
s	Consider two sequences equivalent when one is a substring of another or one sequence contains all words from the other sequence
p	Punctuation insensitive approximate comparison (e.g., separation of sentences with comma, semi-colon or period are considered equivalent)
f	Approximate fingerprint comparison ignoring the order of words in sentences
l	Approximate equivalency when sequences have Levenshtein distance within a configurable threshold (l2 indicates a maximum distance of 2)

Alternative voting and consensus output

Code	Voting and consensus
v	Consensus is reached when there is a single group (set of matching answers) that has the most votes, instead of requiring strict majority vote among all answers
a	Outputs best available answers when consensus is not achieved

n=4

Majority voting requires 3 matching answers $\left\lceil \frac{n}{2} + 1 \right\rceil$

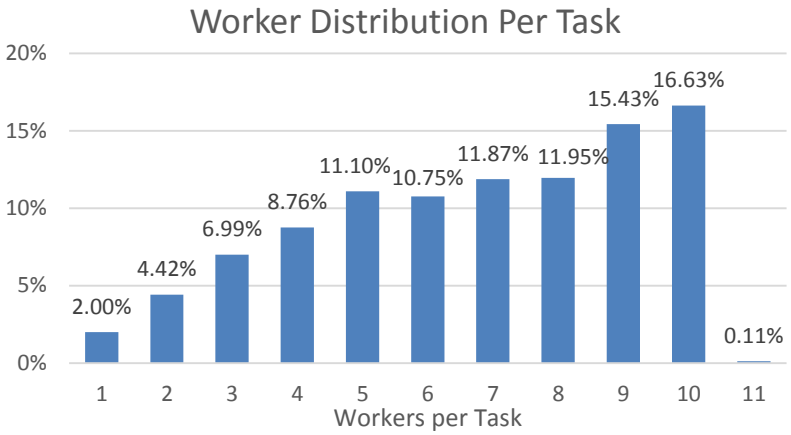
→Consensus not reached

v: Blue set has most votes

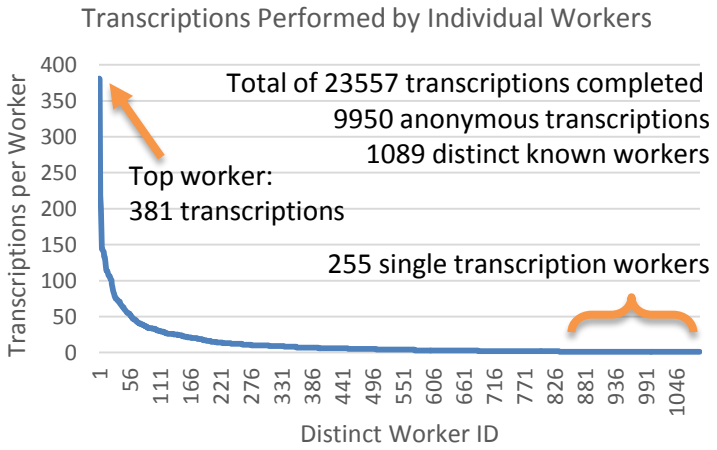
→Consensus reached

Experimental Setup

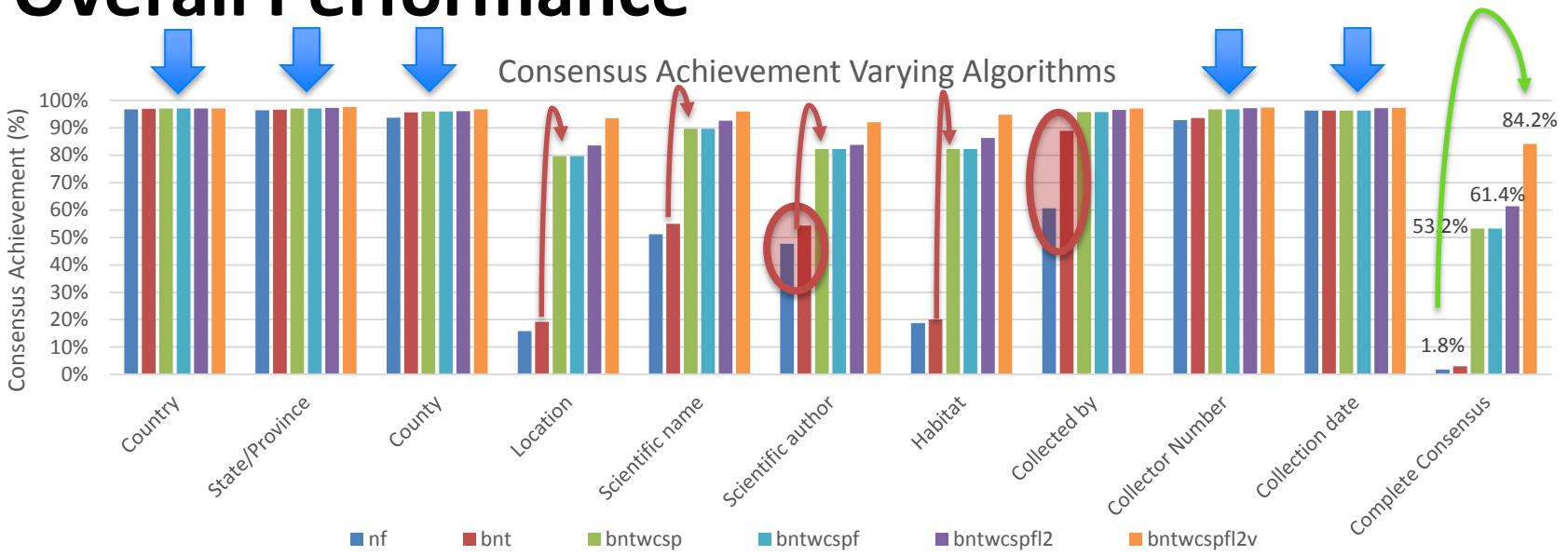
- Notes from Nature
- Herbarium specimens from a single institution
- Configured to require 10 workers per task that yielded close-to-linear distribution due to empty tasks and skips
- 23,557 total transcriptions completed by at least 1,089 distinct workers



Field	Uniq#	Field	Uniq#
Country	39	Location	16,161
State/Province	288	Habitat	15,134
County	655	Collected by	3,380
Scientific name	5,941	Collector Number	3,665
Scientific author	4,088	Collection date	2,287

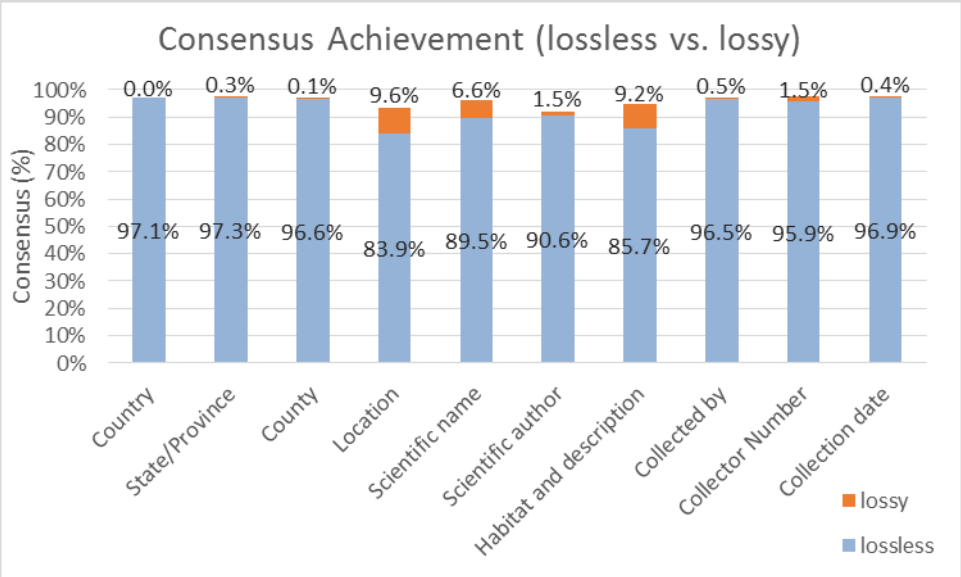


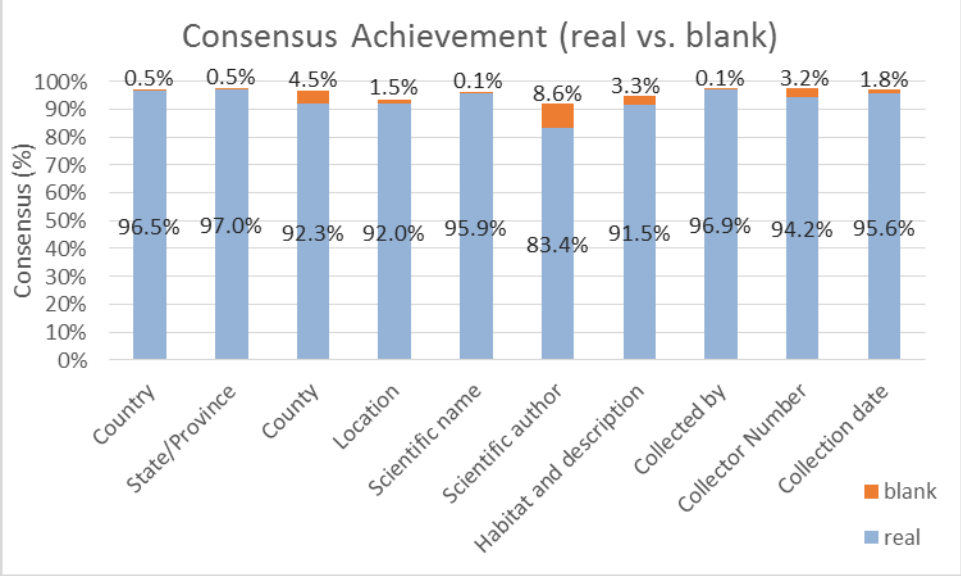
Overall Performance



- Full consensus improvement from 1.8% to 84.2%
- Confirms intuition that *country*, *state/province*, *county*, *collector number* and *collector date* are “easy”
- Lossless algorithms have small impact except for *scientific author* and *collected by*
- Being insensitive to whitespace, punctuation, and letter case as well as considering substrings, provide the greatest improvement when including lossy algorithms in “difficult” fields

Additional Verification



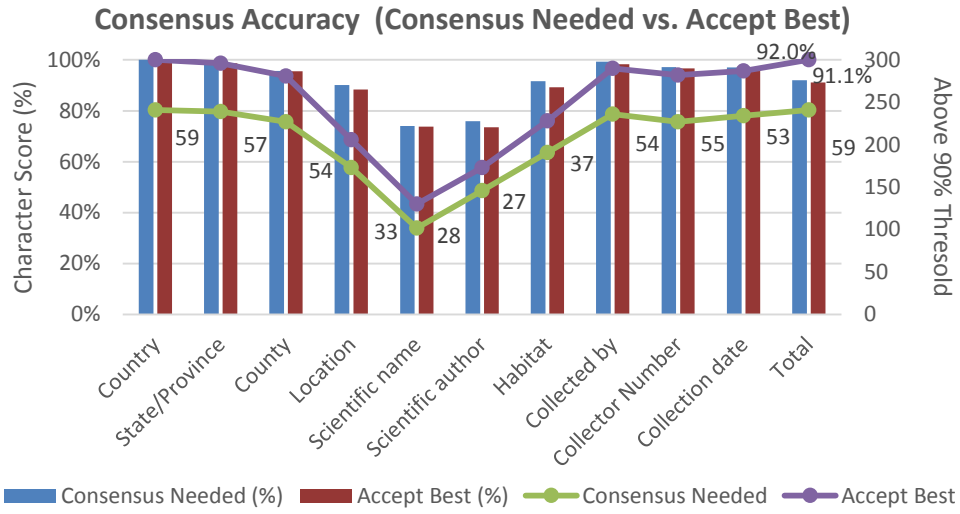


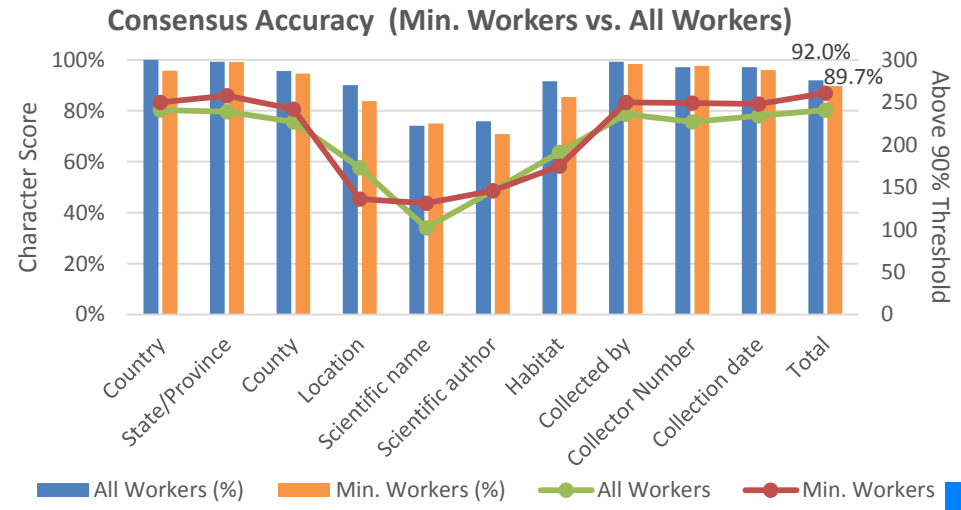
- Consensus reached mainly with lossless algorithms
- Low percentage of blank responses

Consensus Accuracy

$$\frac{2 * matches}{len(s1) + len(s2)}$$

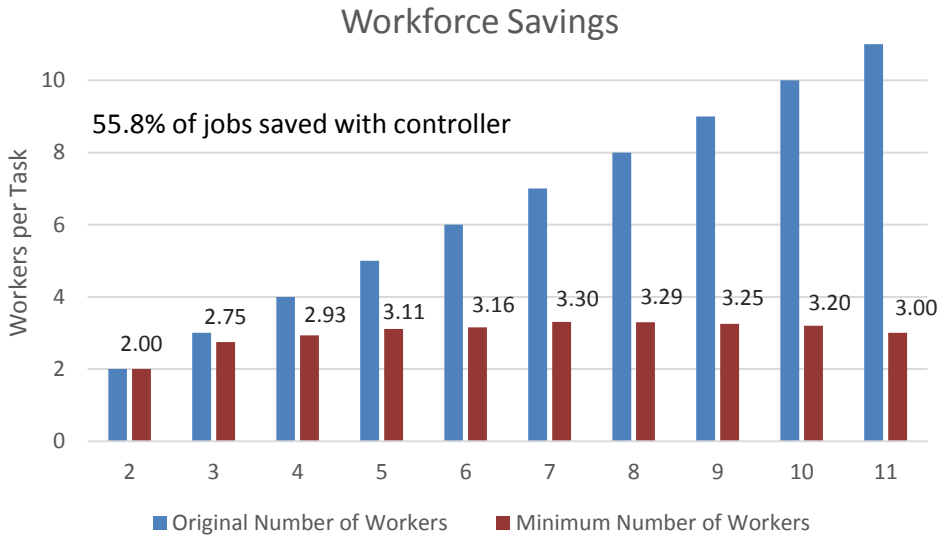
- 300 labels were transcribed by an expert
- Expert had access to information across labels that workers did not have
- Effect on the overall accuracy is minimal
  - 0.9% drop for accepting cases that did not reach consensus
  - 2.3% drop for minimizing the needed workforce





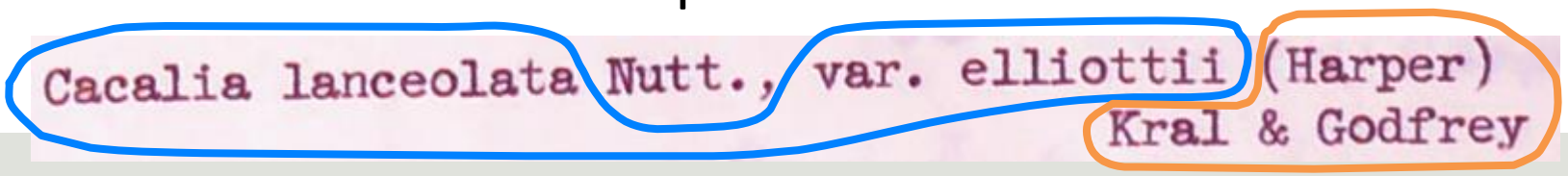
Workforce Savings

- Can be as high as 55.8% for the distribution in the studied dataset
- Good for a fixed setting: 3 workers
- Controller advantage: 3 workers is just a good average, and our results show that there are cases where up to 9 workers were needed to reach overall consensus



Task Design Improvement recommendations

- Restricted user interface improves consensus and accuracy
  - Caution to not restrict valid scenarios (e.g., partial dates, range of dates)
  - Broadly defined fields could be engineered to capture more parsed data (e.g., lat/long, TRS)
- Exploring relationships between tasks
  - Enter collector number and collection date first
  - Update related record to have the same information
- Additional training
  - Problems pronounced in separating scientific names from its authorship



Additional Improvements to Consensus Algorithms

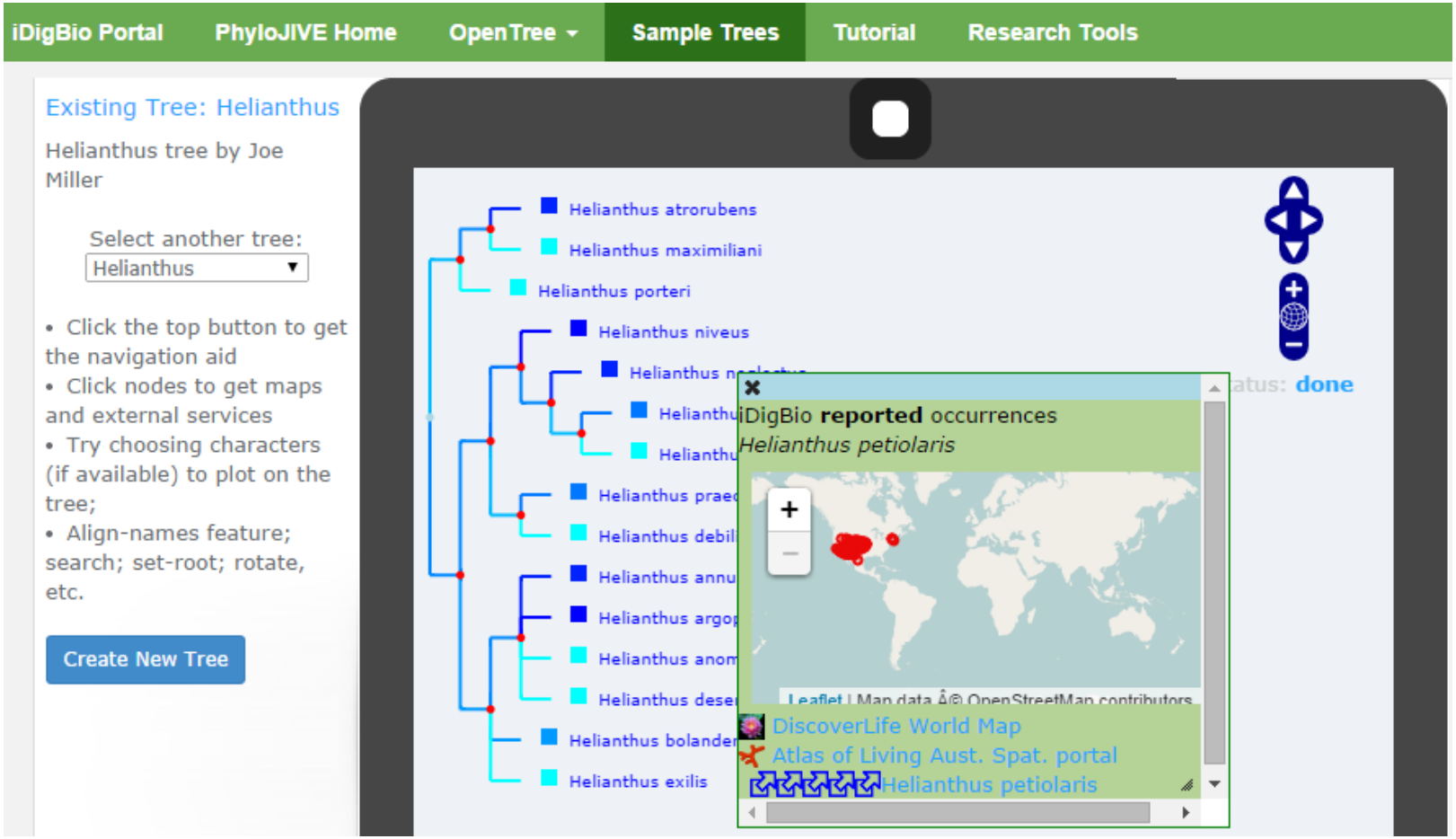
- Code is modular and open; thus, opportunity for:
  - Custom dictionaries could be applied (general dictionaries led to a high number of false positives due to the amount of abbreviation and names)
  - Scientific name parsers
  - External contributions
  - https://github.com/idigbio-citsci-hackathon/CrowdConsensus
- Merge matched outputs after lossy algorithms are applied
  - R. E. Perdue, Jr. and K. Blum
  - Re Perdue Jr and K. Blum
  - R. R. Perdue Jr, K. Blum
- Additional validation across fields (consistency)
- Apply consensus controller on a per-field basis

Recommendations Beyond Crowdsourcing

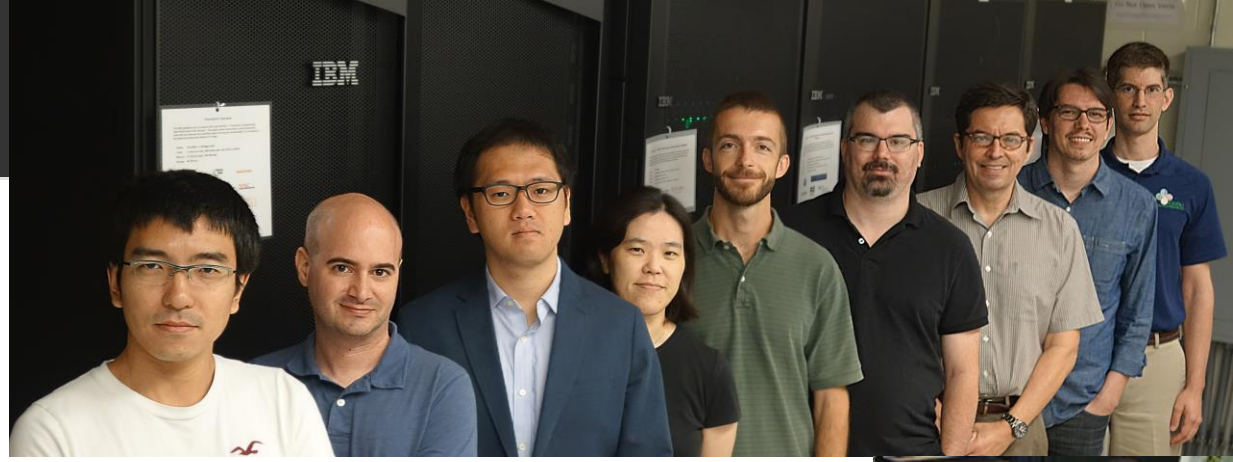
- Leveraging and improving:
  - Optical Character Recognition (OCR)
  - Natural Language Processing (NLP)
- 2-way street scenarios:
  - Use crowdsourcing to select clean text for OCR
  - Use even poor OCR to guide tasks to the right crowd by creating clusters of tasks
  - Use NLP to parse verbatim data from the crowd
  - Improve NLP and OCR training with additional data from the crowd

Additional parsed data

- PhiloJIVE: http://phylojive.acis.ufl.edu



iDigBio
Integrated Digitized Biocollections



Questions?

www.idigbio.org

facebook.com/iDigBio

twitter.com/iDigBio

vimeo.com/idigbio

idigbio.org/rss-feed.xml

webcal://www.idigbio.org/events-calendar/export.ics

Thank you!

iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.