Herbarium Labels Transcription Crowdsourcing Consensus

Joshua Campbell,
Andréa Matsunaga,
José A.B. Fortes

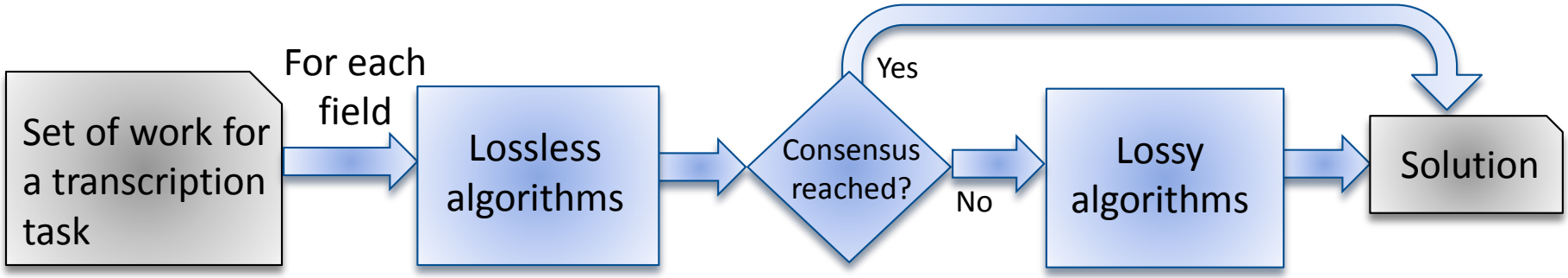
Supported by NSF Award EF-1115210

iDigBio Transcription Hackathon
Gainesville, USA
Monday 16th - Friday 20th of December 2013

Transcription Task

- What is being transcribed:
  - Country
  - State/Province
  - County
  - Scientific name
  - Scientific author
  - Location
  - Habitat and description
  - Collected by
  - Collector Number
  - Collection date

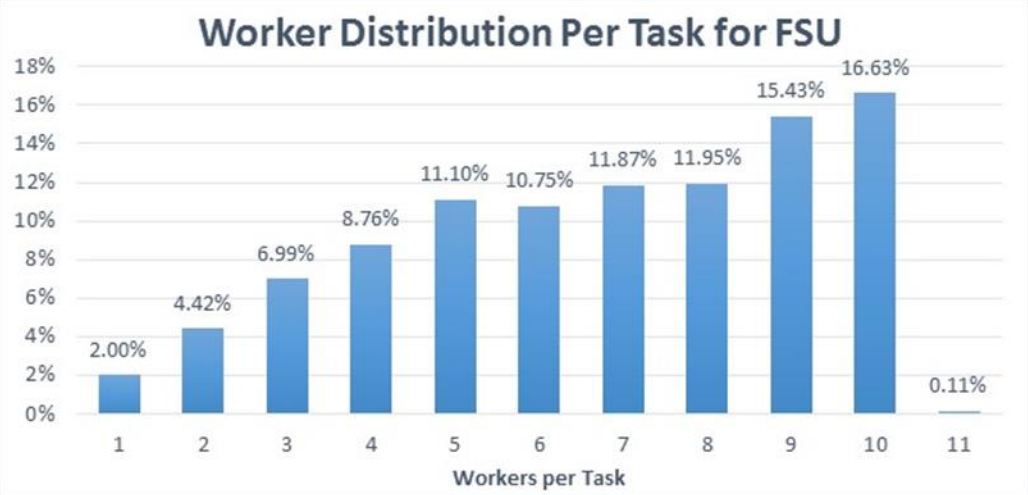
Consensus Approach

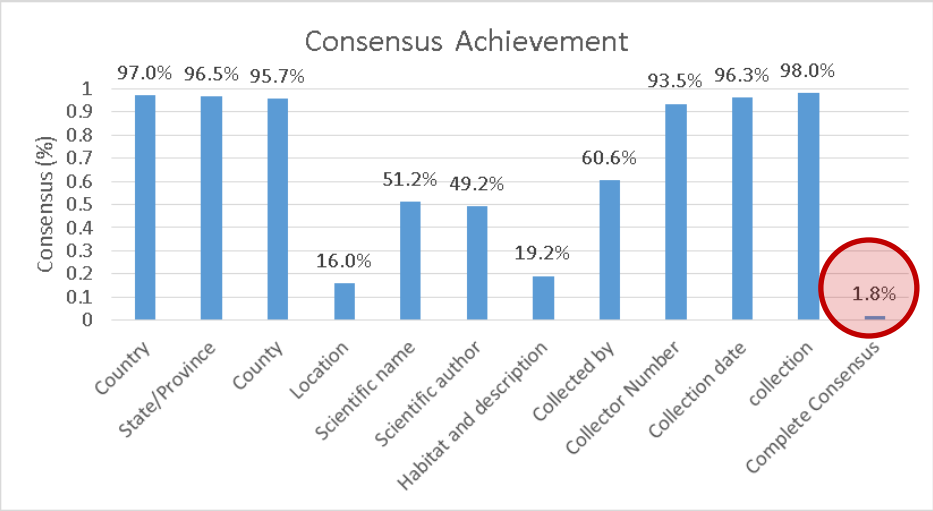


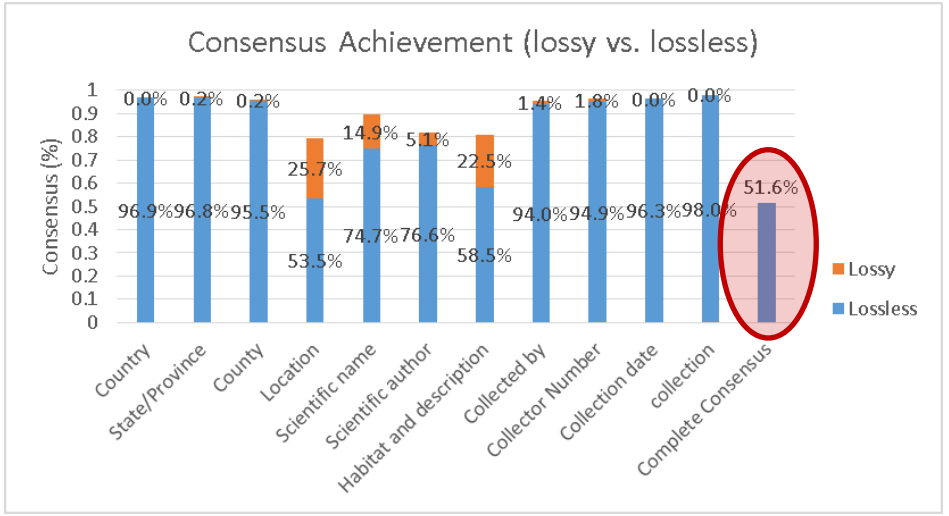
- Lossless:
  - Extra whitespace removal
  - Normalization of blank responses
  - Normalization of abbreviations (hwy, hiway, highway)
  - Normalize TRS
  - Normalize special characters (½, ¼, º)
  - Normalize proper names (A. B. vs. A.B. vs. AB)
  - Spell correction
- Lossy:
  - Ignore letter case → prefer title case for names, sentence case otherwise
  - Ignore punctuation → prefer extra puntuation
  - Ignore stop words → prefer presence of stop words
  - Ignore diacritics (á, ê, õ, ü) → prefer version with diacritics
  - Allow differences up to a certain edit distance

Basically, the same concepts used by OpenRefine

Preliminary Results & Goals







- Continue improving the consensus methods, tune number of citizen scientists needed for a task, minimize future crowdsourcing work