Task Design and Crowd Sentiment in Biocollections Information Extraction

Ícaro Alzuru, Andréa Matsunaga, Maurício Tsugawa, and José A.B. Fortes

Advanced Computing and Information Systems (ACIS) Laboratory

University of Florida, Gainesville, USA

3rd IEEE International Conference on Collaboration and Internet Computing

October 16th, 2017

San Jose, California

HuMaIN is funded by a grant from the National Science Foundation's ACI Division of Advanced Cyberinfrastructure (Award Number: 1535086). 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.

ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

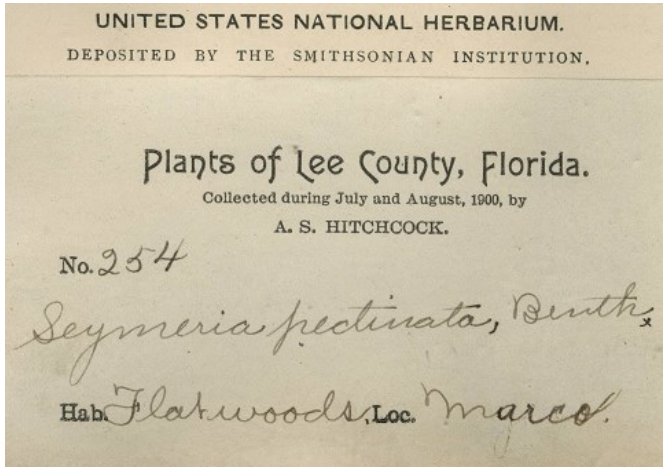
Agenda

- Biocollections
- HuMaIN project
- Current Information Extraction (IE) interfaces in Biocollections

Biological Collections

Photo by Chip Clark. Bird Collection, Department of Vertebrate Zoology, Smithsonian Institution's National Museum of Natural History. In the foreground is Roxie Laybourne, a feather identification expert.

- For about 250 years humans have been collecting biological material. The metadata from biocollections can be used to study pests, biodiversity, climate change, species invasions, historical natural disasters, diseases, and other environmental issues. [1]
- It has been estimated in 1 billion the specimens in the USA which information could be digitized [1], and 3 billion in the whole world [2].
- In USA, since 2012, iDigBio has aggregated more than 105 M. digitized records [3]. Worldwide, GBIF accumulates more than 740 M. records in its database and website. [4]
- The extraction of the metadata is a difficult task that requires humans.

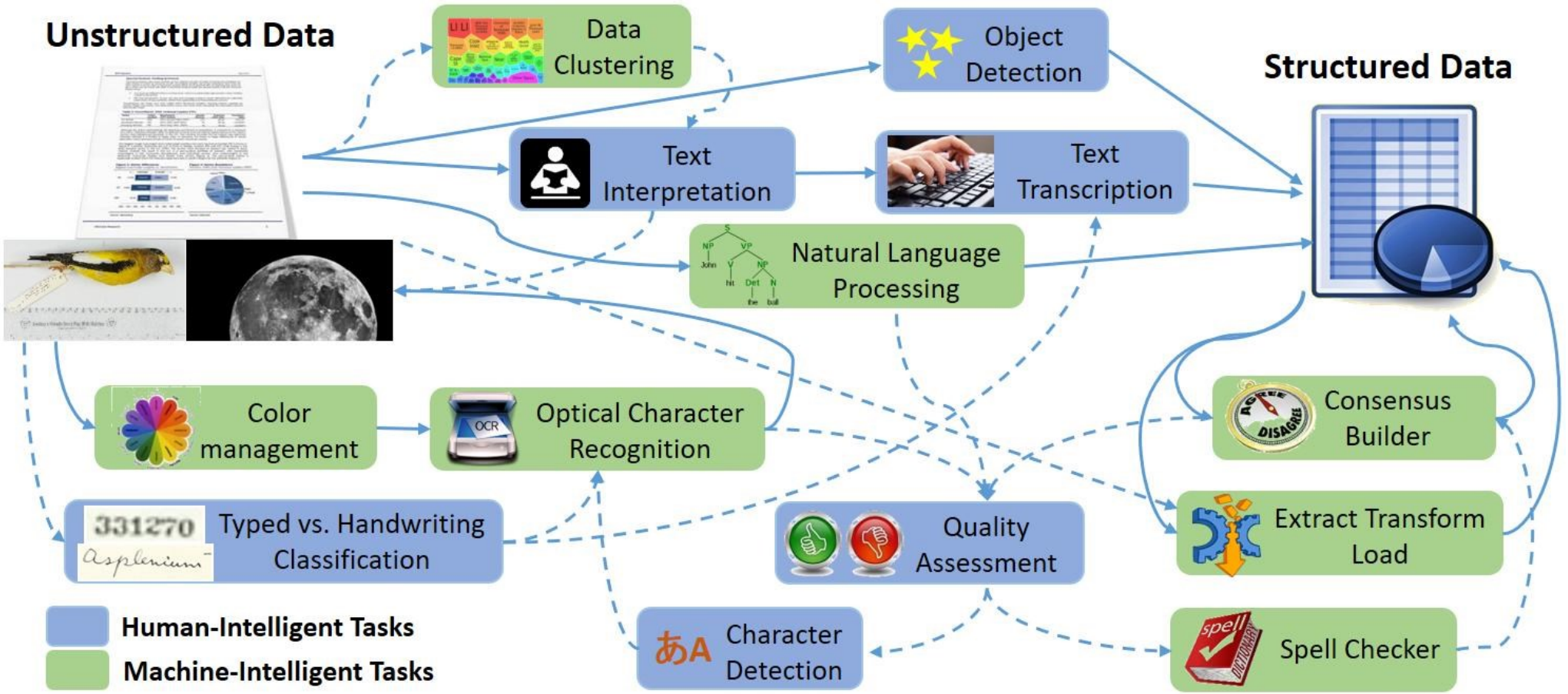


ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

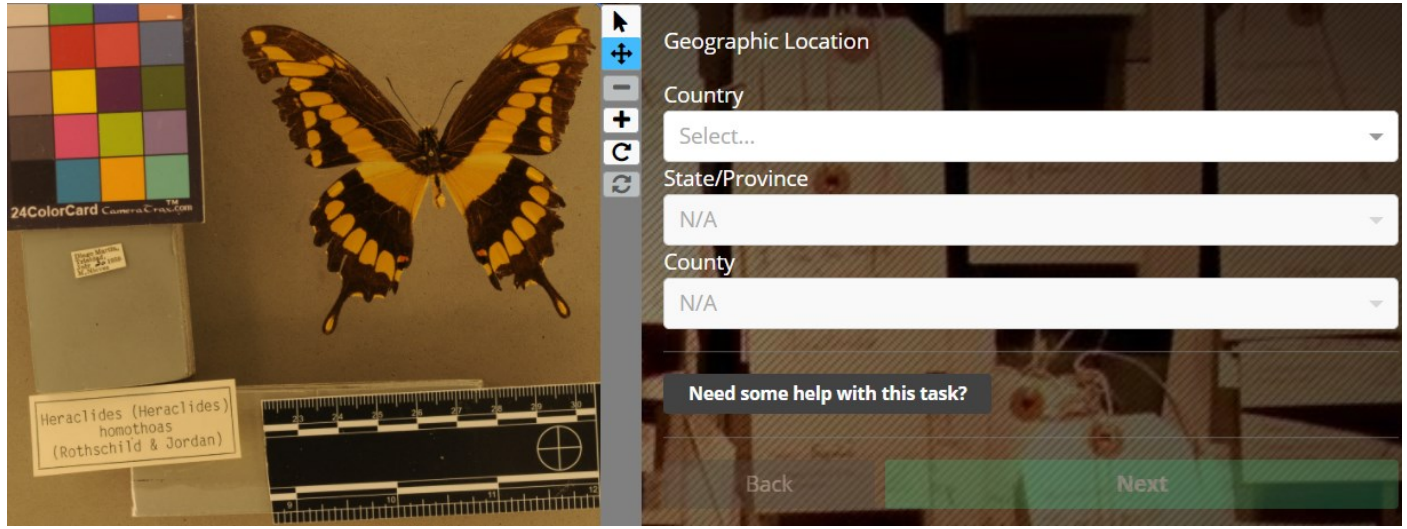
HuMaIN

Human and Machine Intelligent Software Elements for Cost-Effective Scientific Data Digitization

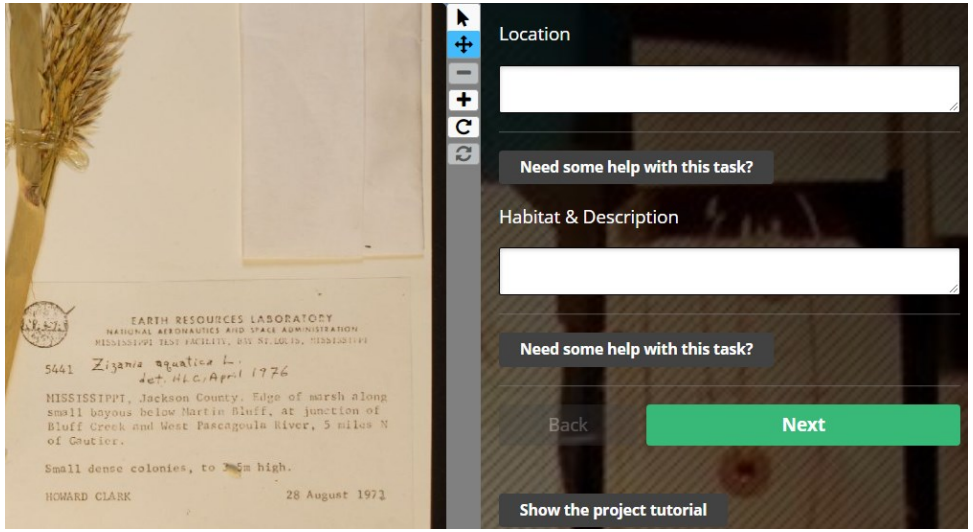


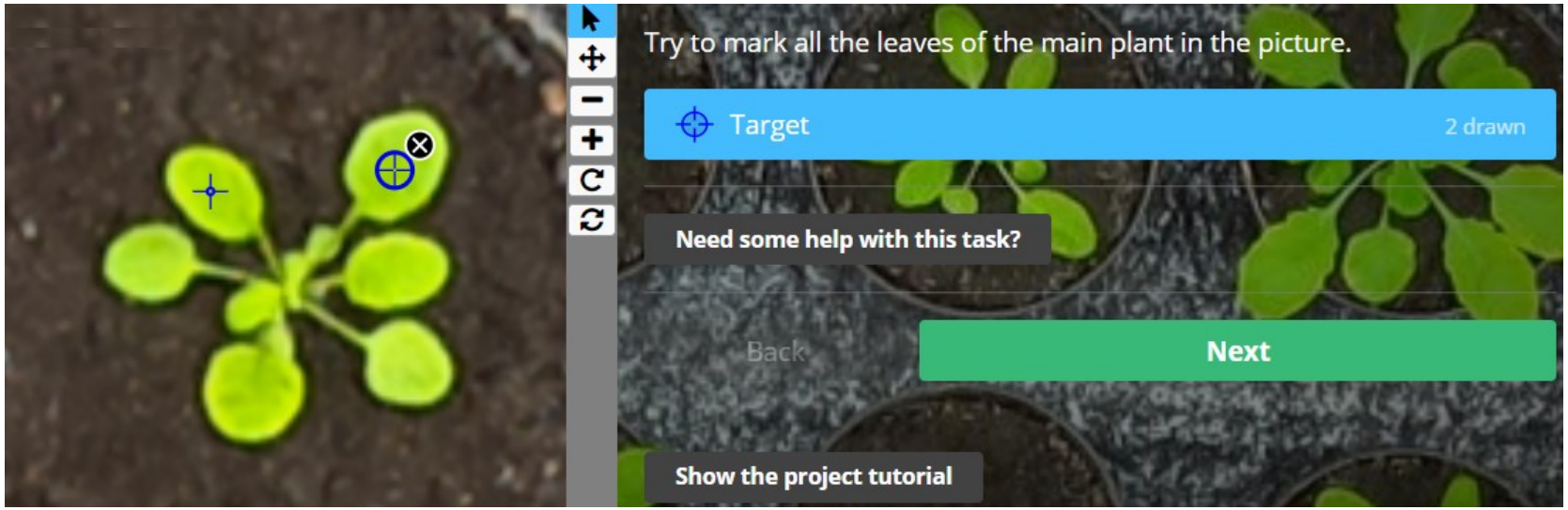
IE Interfaces for Biocollections

Notes from Nature - Select values from a list of options



Notes from Nature - Transcribe (type)



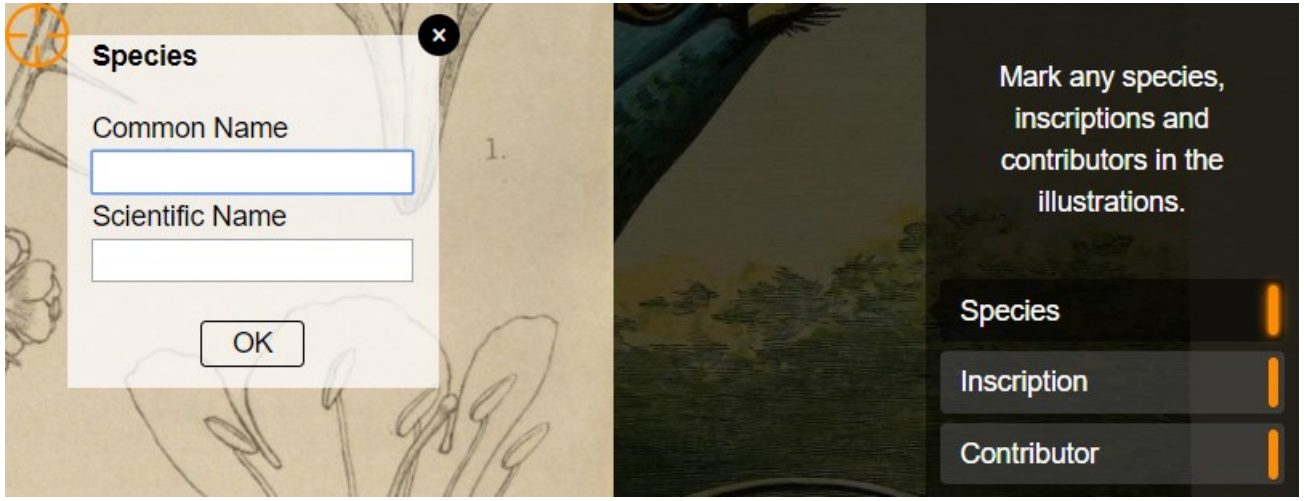


Zooniverse - Mark

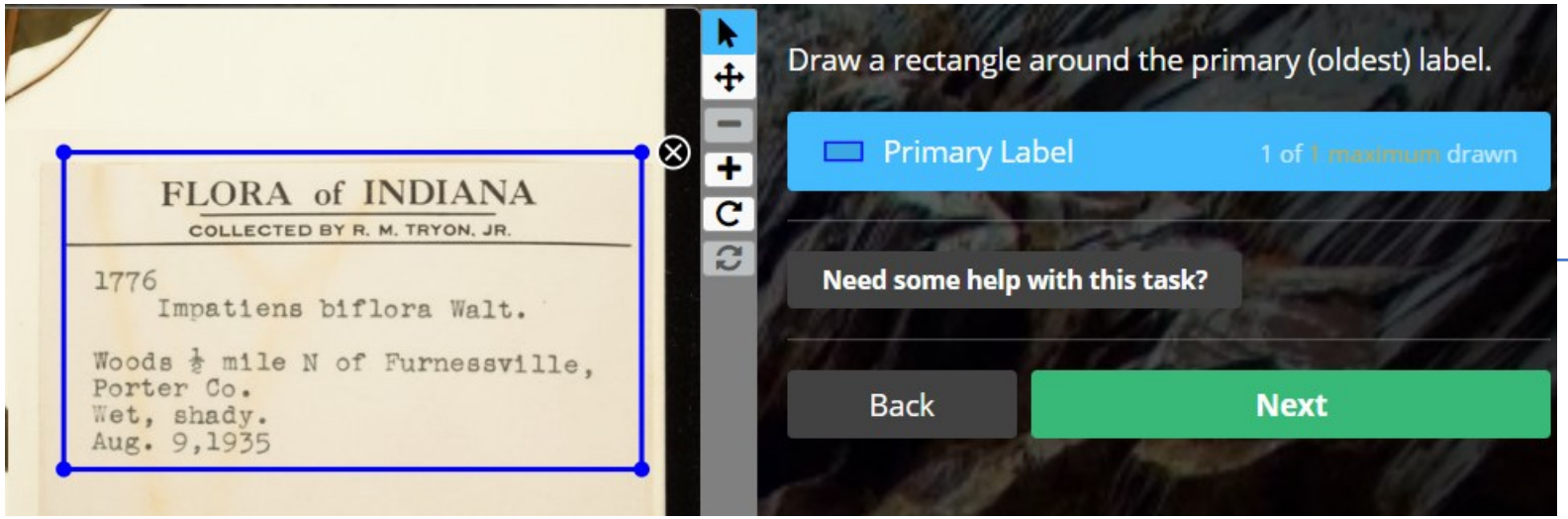
ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

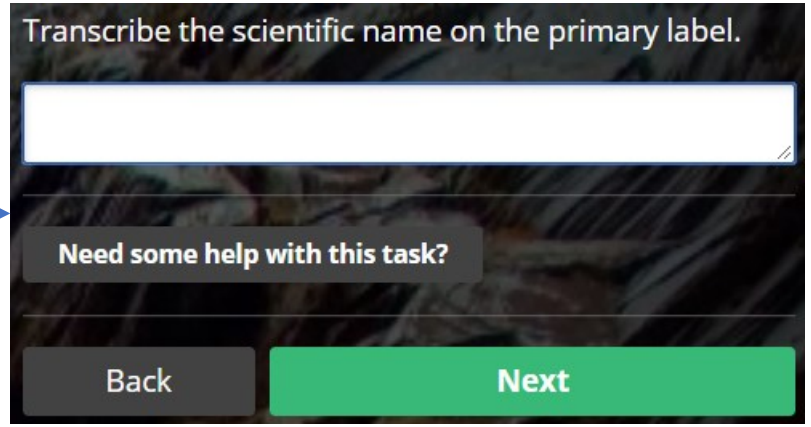
UF UNIVERSITY of FLORIDA

IE Interfaces for Biocollections



Science Gossip: Mark + Transcribe
(as many items you find in an image)





Zooniverse – Label? (Y/N) + Delimit + Transcribe

The problem -> The study

- At present, biocollections' IE is based on crowdsourcing.
- The most commonly used interface interactions to enter information are:
  - Transcription
  - Selection (lists, checkboxes)
  - Other mouse interactions (mark, drag)
- **Does any of these interfaces provide an advantage on duration or quality of the results over the others?**
- Some crowdsourcing apps request the information by field, others ask to complete several fields at once.
- **How task granularity and these different interface options impact output quality and processing time?**
- **What is the opinion of the crowd about these alternatives?**

Related Work

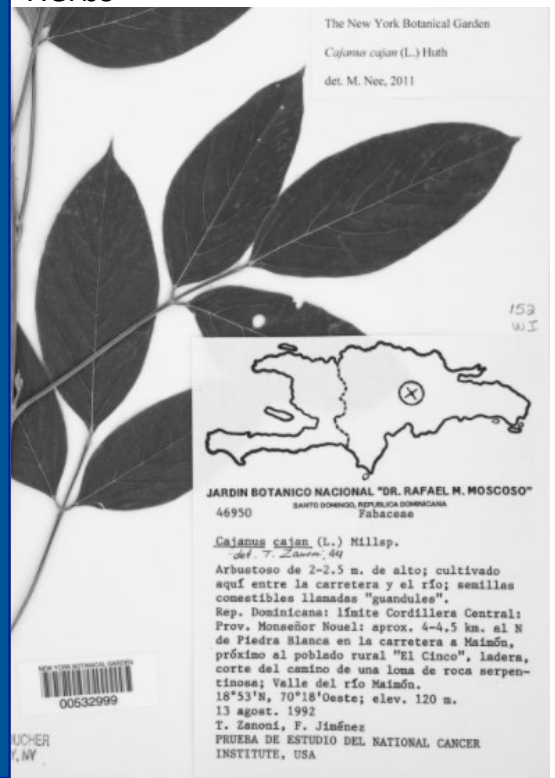
- State of art in biocollections' IE interfaces and good practices:
  - *More general, platform specific, quality of image, tutorial, clear objective.*
- Microtasks vs. Macrotasks (granularity):
  - *Microtasks generate better quality. General purpose crowdsourcing.*
- Gamification, competitiveness, reward, and other engagement strategies:
  - *Highlight the importance of keeping volunteers engaged.*
- Human-Computer Interaction, geometrical factors, and interface objects in task efficiency.
- Quality oriented papers:
  - *Cost, duration, and crowd are usually forgotten.*

Experimental Design (1/3)

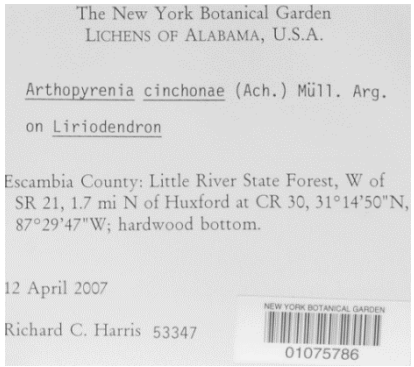
Dataset [5]:

- Three different collections: Insects, Herbs, and Lichens (400 images).
- Subset of 100 images (34, 33, 33)

Herbs



Lichens



Insects

30 tasks were used throughout this study:

- *Transcription* of:
  - 12 fields: *Event date*, *Scientific name*, *Identified by*, *Country*, *State*, *County*, *Latitude*, *Longitude*, *Elevation*, *Locality*, *Habitat*, and *Recorded by*.
  - 8 fields (textual): *Scientific name*, *Identified by*, *Country*, *State*, *County*, *Locality*, *Habitat*, and *Recorded by*.
  - 4 fields (numerical): *Event date*, *Latitude*, *Longitude*, *Elevation*.
  - Each of the 12 fields, independently.
- *Selection* of:
  - *Event date*.
  - *Identified by*.
  - *Country*, *State*, and *County*.
- *Cropping* of:
  - Each of the 12 fields.

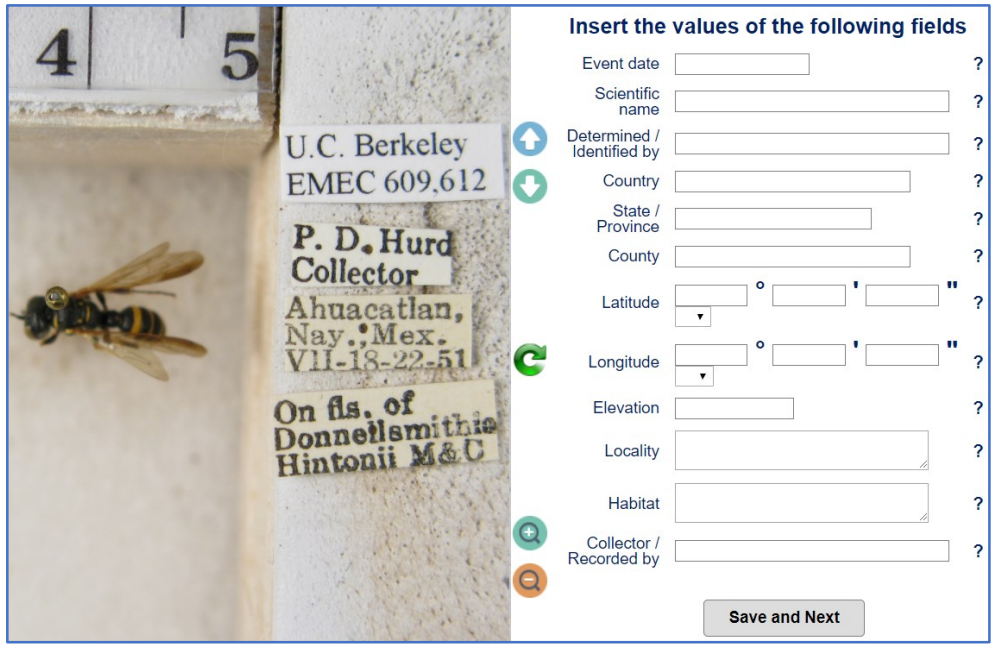
UF UNIVERSITY of FLORIDA

Experimental Design (2/3)

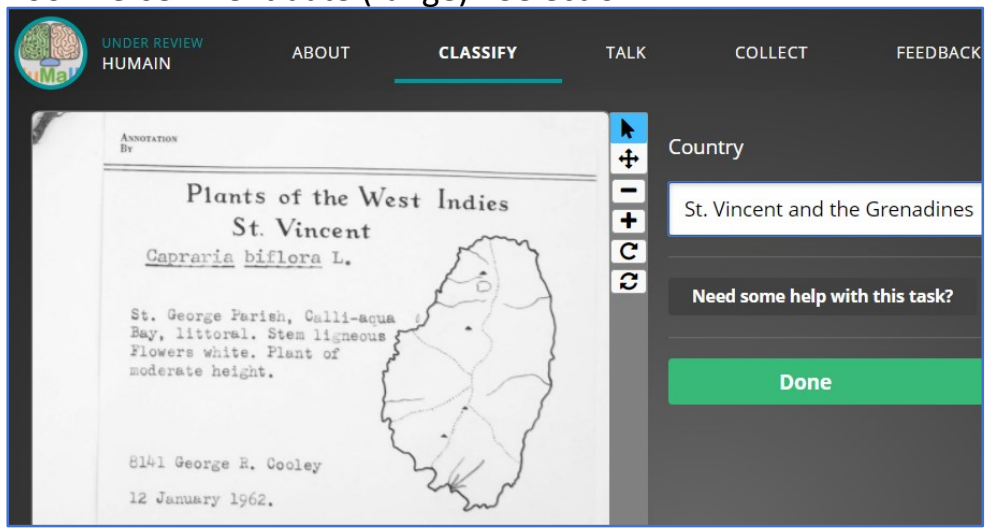
Web platforms:

- HuMaIN (on-site): 41 participants.
  - They were paid $10/hour
- Zooniverse: 436 users.
  - Only Transcription

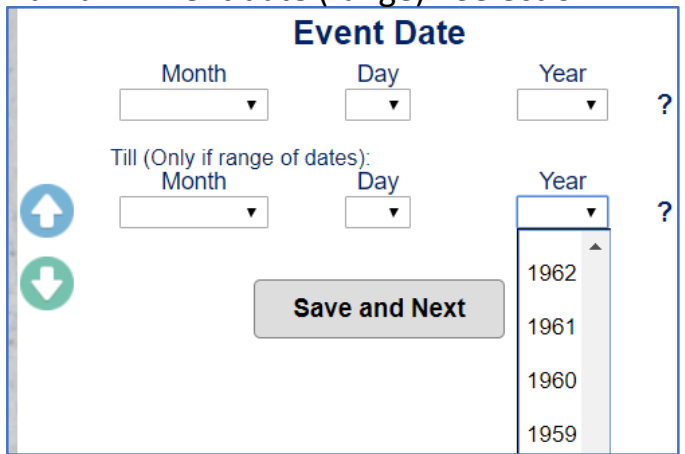
HuMaIN:
12 Fields
Transcription



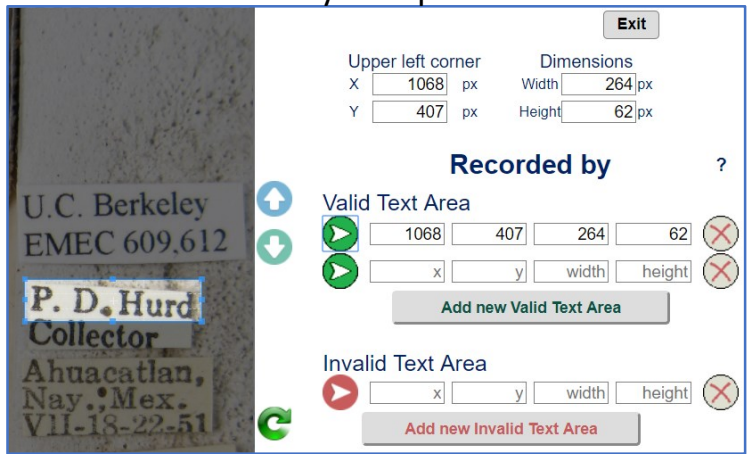
Zooniverse: Event date (range) - Selection



HuMaIN: Event date (range) - Selection



HuMaIN: Recorded by - Crop



ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

Experimental Design (3/3)

Computation of Quality

Strings were compared using the **Damerau-Levenshtein** algorithm (minimum amount of insertions, deletions, substitutions, and transpositions of two adjacent characters, required to convert one string into the other) to generate a **similarity** value:

$$sim_{DL}(x, y) = 1 - \frac{DL\ distance(x,y)}{\max(|x|,|y|)}$$

0 -> Totally different strings

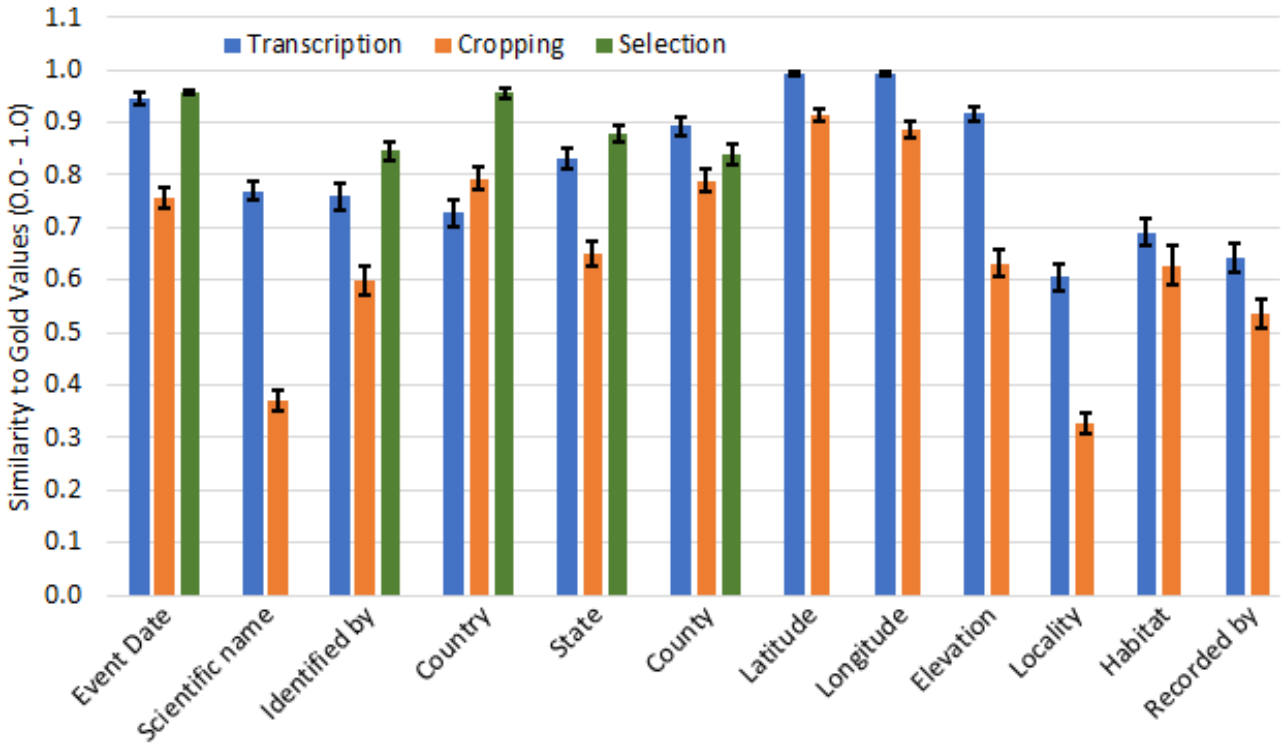
1 -> Identical strings

Extracted Values are categorized using the confusion matrix terminology:

- TP: correctly identified value. Quality is estimated using the DL similarity.
- FN: incorrect omitted value. Quality = 0.
- FP: incorrectly omitted value. Quality = 0.
- TN: correctly omitted value. Quality = 1.

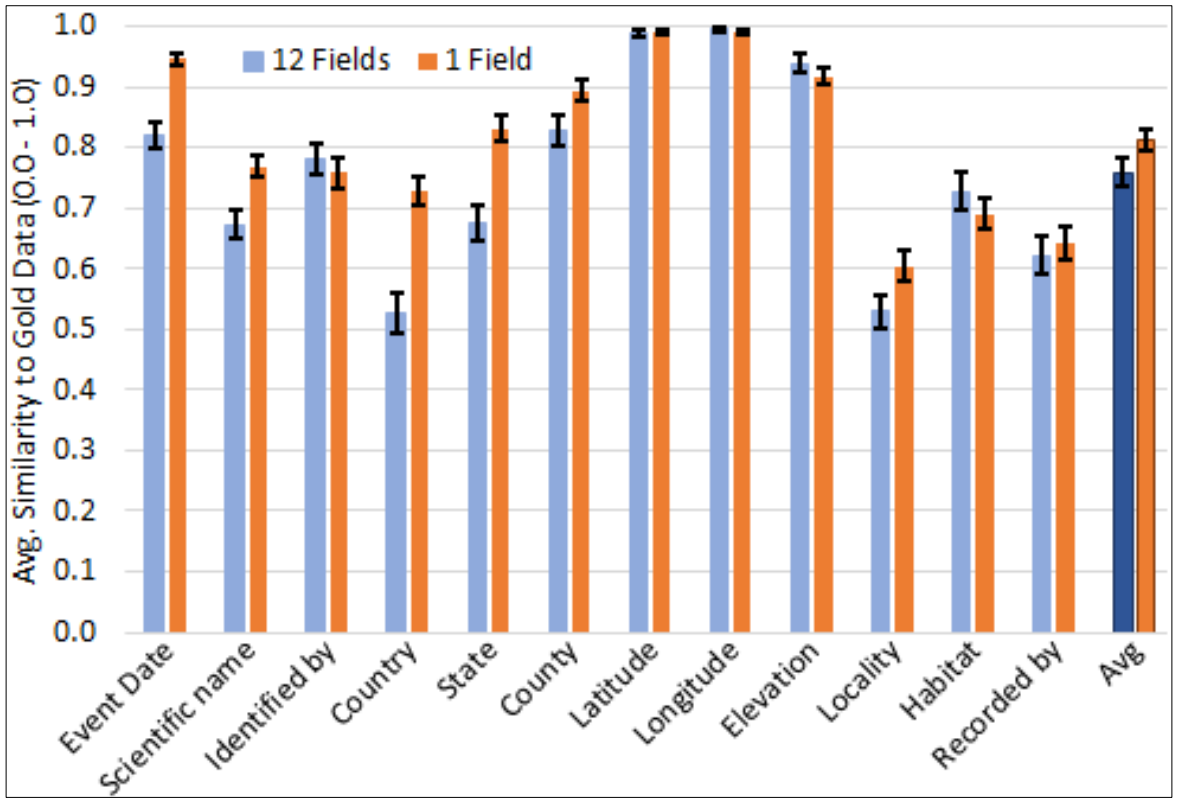
Results - Quality by Interface Type and Field

Similarity (Quality) of extracted values when compared to the gold (experts') output.

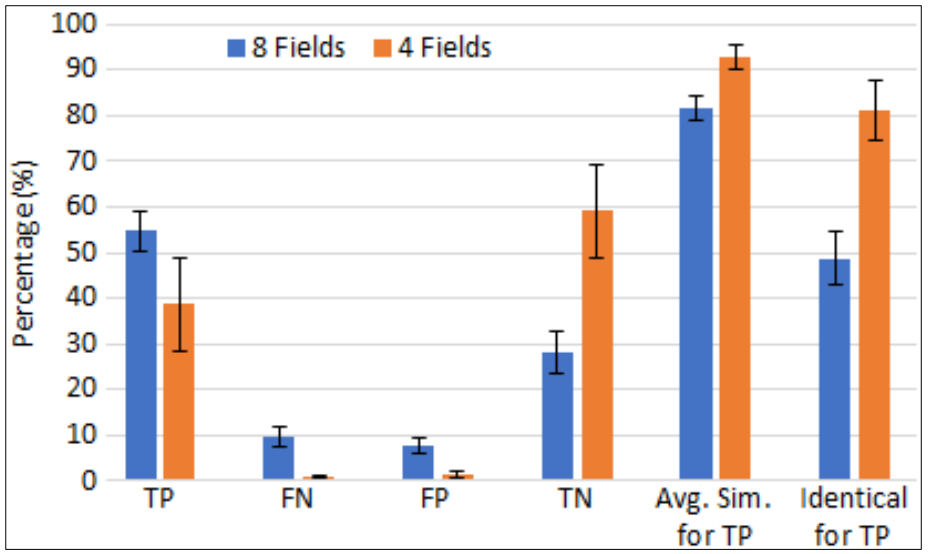


- *Selection* generated a result of higher quality than *Transcription*, with the exception of *Country*.
- *Cropping* + OCR generated the results with the worst quality. But it depends on:
  - the quality of the images
  - the quality of the OCR software and how trained it is to recognize text in similar conditions.
- Two users negatively affected the quality of *Country*'s output for *Selection* because they inferred non existent country values.

Results – Quality by Granularity

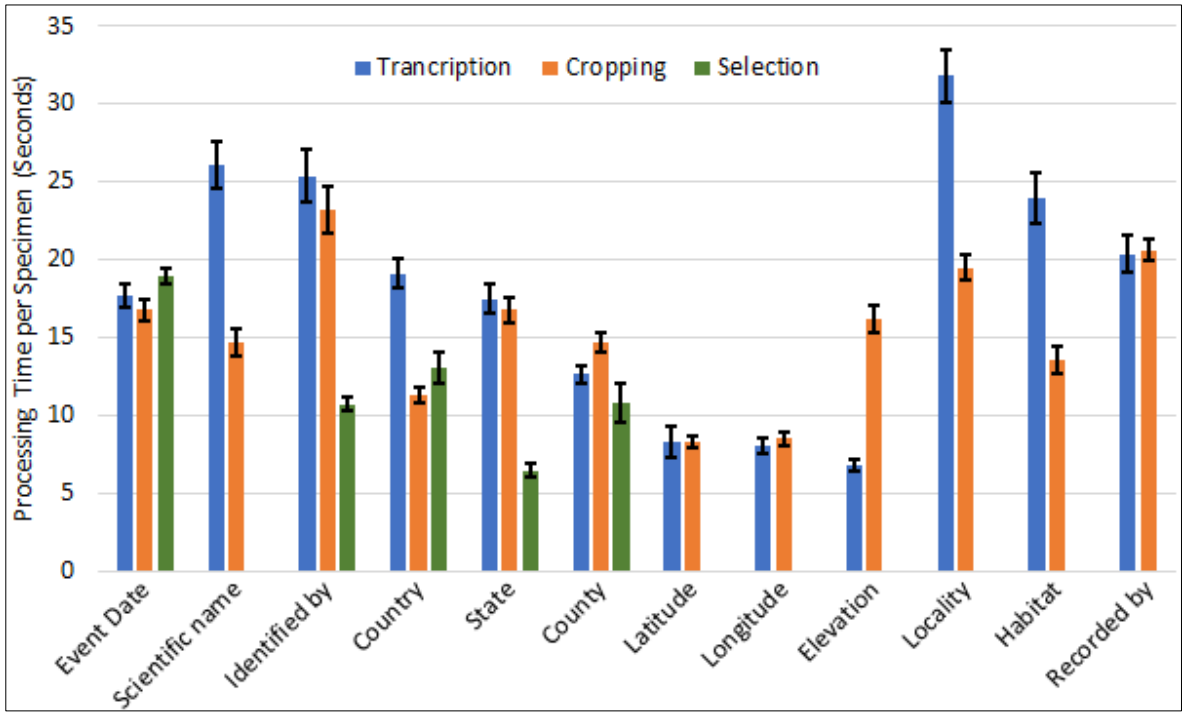


- Single field tasks improved the overall quality of the result by 7.25%.



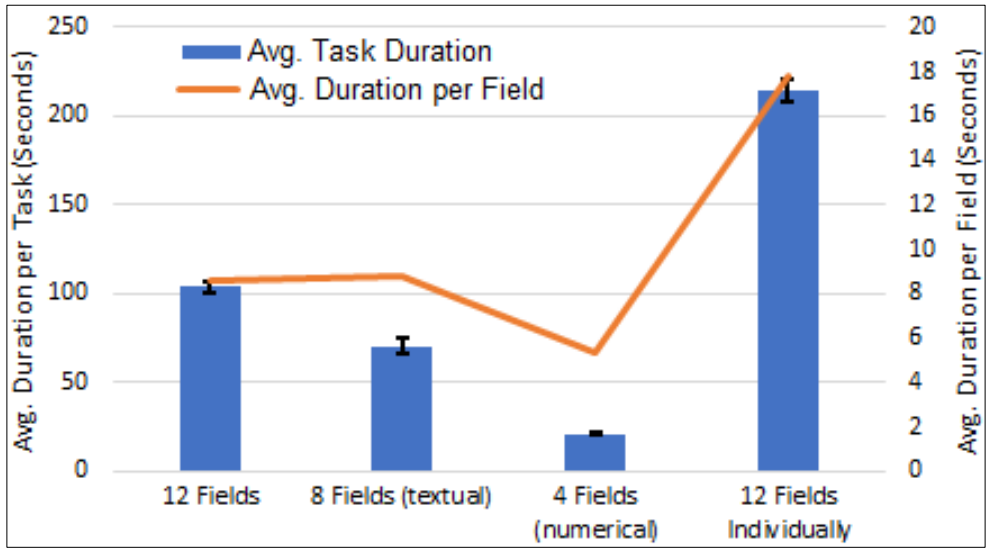
- Numerical fields generated results with 11% higher similarity and 33% more identical values than textual fields.

Results - Duration by Interface Type and Field



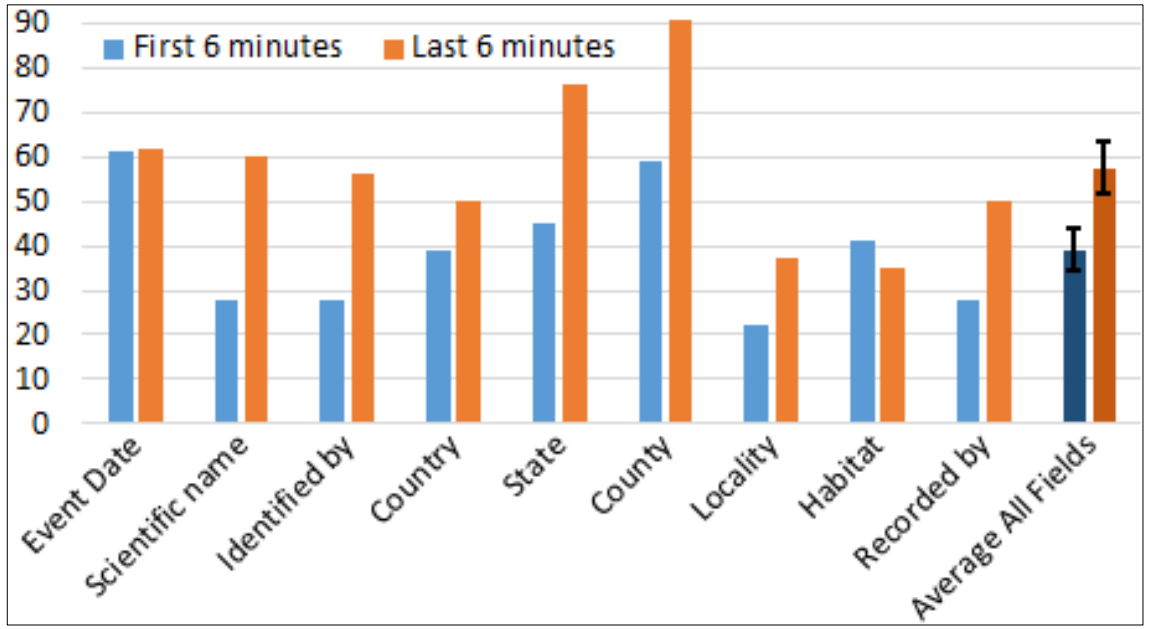
- *Selection* was faster than *Transcription* and *Cropping* in 3 of the 5 fields.
- In *Event date*, users have to select 3 values, for the most common case.
- In fields that require long text, such as *Scientific name*, *Locality*, and *Habitat*; *Transcription* becomes a slow option in comparison to the other two options.
- *Selection* has the advantage that normalizes the output values, but its it cannot always be implemented.

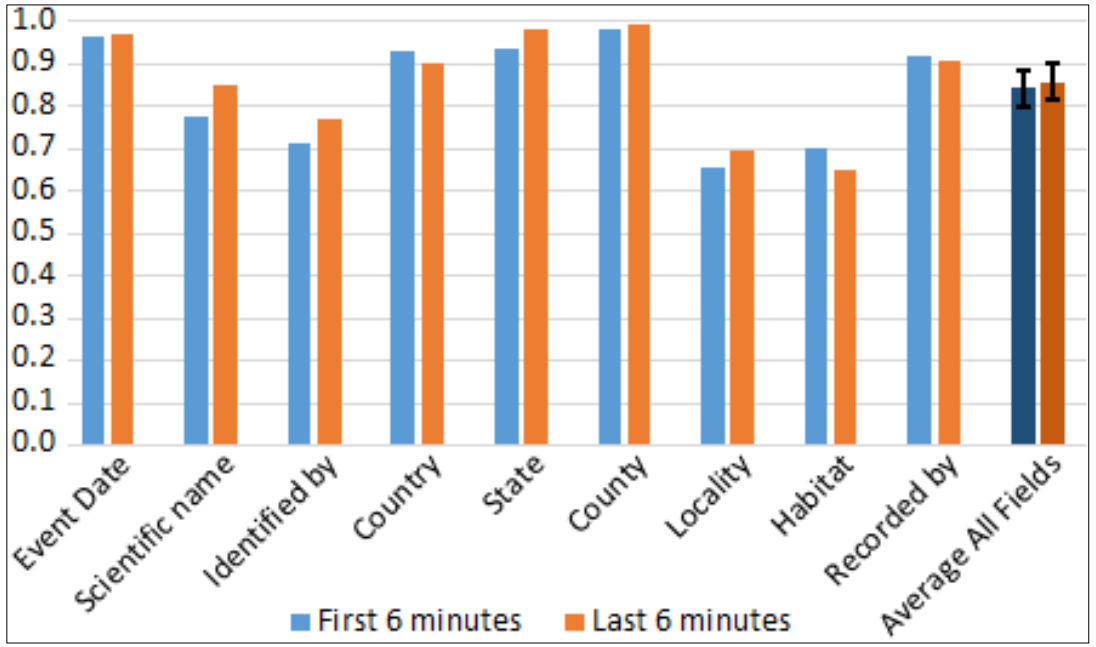
Results – Duration by Granularity



- 12 single field tasks takes twice the time taken to complete the 12 fields compound task (104 vs. 208 seconds).
- Textual fields take more time to be transcribed than numerical fields.

Results – Learning Process





- With the exception of *Habitat*, users have a higher rate of processed images towards the end of their work session.
- Users require some time or practice to internalize the concept, learn how to identify the value in the image and use the interface.

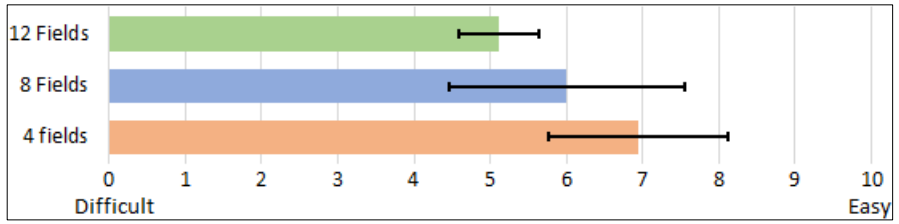
- However, this does not hold true for the output quality, which basically stays the same at the beginning and towards the end of the experiments.

ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

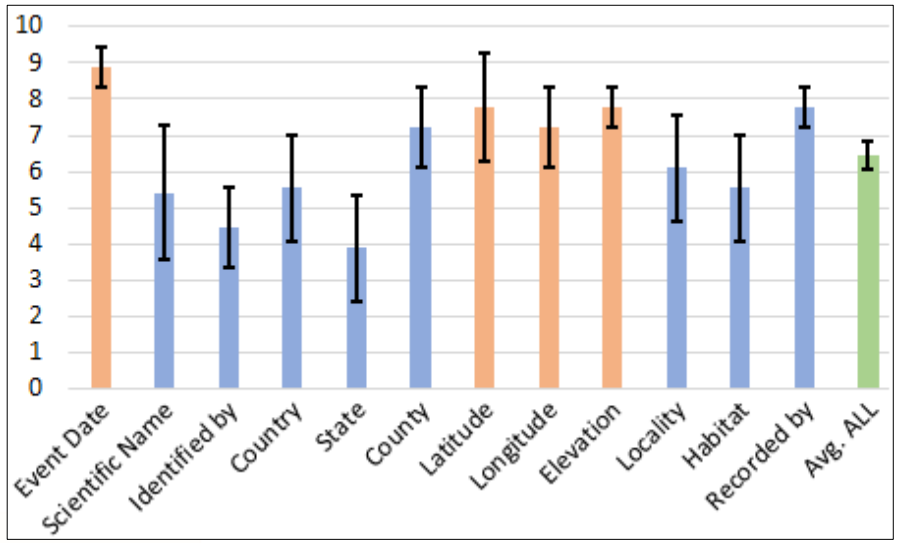
Results – Crowd Sentiment (1/2)

The experiment was perceived as **slightly easy**

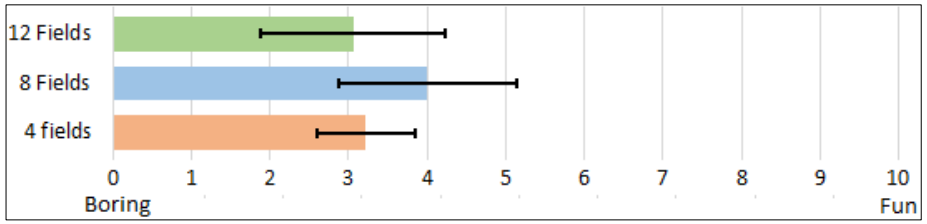


Numerical fields are **easier** to complete than textual fields.

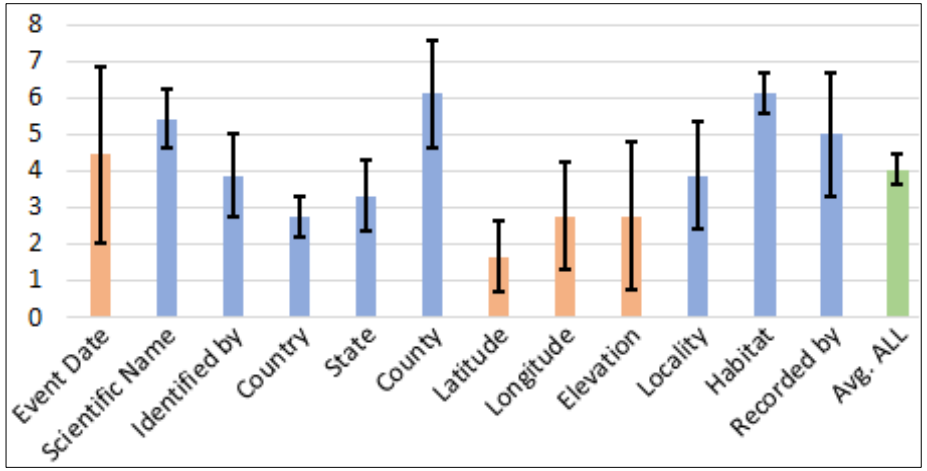
State was difficult because there were specimens from several countries.



The experiment was perceived as **boring**



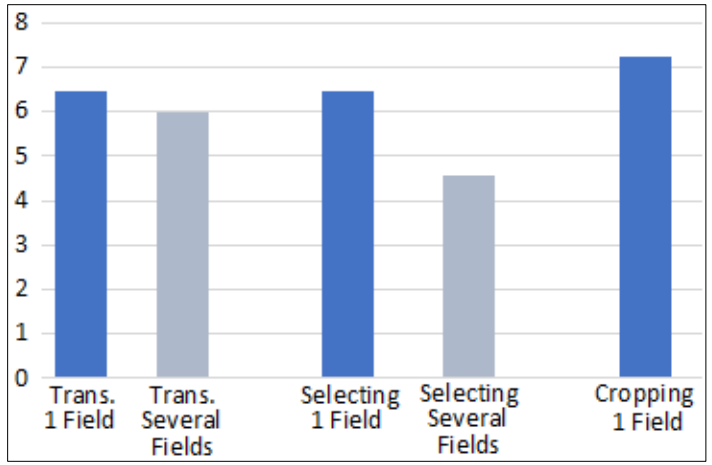
Numerical fields are **more boring** to complete than textual fields.



ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

Results – Crowd Sentiment (2/2)



Conclusions

- Selection generates higher quality outputs than Transcription.

Thank you!

Any question?

HuMaIN is funded by a grant from the National Science Foundation's ACI Division of Advanced Cyberinfrastructure (Award Number: 1535086). 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.

ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

References

1. J. Hanken, “Biodiversity online: toward a network integrated biocollections alliance,” Bioscience, vol. 63, pp. 789-790, 2013.
2. A.H. Ariño, “Approaches to estimating the universe of natural history collections data,” Biodiversity Informatics, vol. 7, 2010.
3. Integrated Digitized Biocollections (iDigBio). [Online]. Available: https://www.idigbio.org/. [Accessed: 07-Jul-2017]
4. Global Biodiversity Information Facility. [Online]. Available: http://www.gbif.org/. [Accessed: 07-Jul-2017]
5. Label-data. [Online]. Available: https://github.com/idigbio-aocr/label-data/. [Accessed: 01-Oct-2017]