# Enabling the TCNs and Collaborators
# Breakout Group #4: Label Capture & Post-Processing

**Facilitator Name:** Jim Hanken

**Scribe Name:** Grant Godden

**Time Allotted:** 150 minutes

**Group Participant List:** Corinna Gries, Umberto Ravaioli, Robert Naczi, Alan Prather, James Macklin, Deb Paul, Pam Soltis, Renato Figueiredo, Shari Ellis

**Objectives:**
Discuss and produce a report to summarize label capture within the ADBC community. Focus on opportunities to leverage existing tools/systems, standards, practices and techniques. Nominate a reporter to deliver a 15-minute summary report to the plenary session at the conclusion of your session.

**Deliverables:**

1. Define and order at least five critical challenges faced by the TCNs related to label capture and post-processing (#1 is the most critical challenge).

| Rank Order | Challenges Related to Label Capture and Post-Processing |
|---|---|
| 1 | Capture of data on handwritten labels |
| 2 | Challenges parsing data into fields from OCR results (i.e., Natural language processing) |
| 3 | Overcoming unique challenges with image processing and image manipulation |
| 4 | Total data capture: i.e., to accommodate multiple images and multiple specimen labels from a single specimen |
| 5 | Providing cost-effective technologies for automated workflows |
| 6 | Reconciling differences in identification histories (i.e., variable specimen annotations) |
| 7 | Bridging a disconnect between “filed as” versus “specimen label data” |

2. Identify and order up to five existing practices and techniques that can be leveraged for label capture and post-processing (#1 is the most preferred practice/technique). If more than five, focus on the five that are currently the most viable, commonplace, and applicable to the needs of the TCNs and collaborators, while keeping a list of all references to existing practices.

| **Rank Order** | **Label Capture and Post-Processing Practices and Techniques** |
|---|---|
| 1 | Borrow from existing commercial/proprietary software and practices used by industry/government (CAPTCHA) |
| 2 | Imaging (pen camera, gigapan, etc.) |
| 3 | Image enhancement and manipulation |
| 4 | Multi-keying (crowd-sourcing) |

3. Identify and order up to five existing standards that can be leveraged for label capture and post-processing. If more than five, focus on the five that are currently the most viable, commonplace, and applicable to the needs of the TCNs and collaborators. Explain the choices.

| **Rank Order** | **Label Capture/Processing Standards** | **Explanation of Selections** |
|---|---|---|
| 1 | DARWINCORE/ABCD | |
| 2 | AUDUBONCORE | |
| 3 | OGC: Open Geospatial Consortium | |
| 4 | Controlled taxon name authority (ITIS, CoL, etc.) | |
| 5 | ISO Standard (Geography) | |
| Other (non-prioritized) | Other controlled names (phenology, etc.) | |

4. Identify and order up to five existing tools/systems that can be leveraged for label capture and post-processing (#1 is the most preferred tool/system). If more than five are proposed, focus on the five that are currently the most viable and beneficial to the greatest number of stakeholders. Explain the choices. Link tools/systems to the practices/techniques (identified in Deliverable #2) and standards (identified in Deliverable #3) that each enables or supports.

| Rank Order | Label Capture and Post-Processing Tools | Explanation of Selections | Linked Practices/ Techniques (Line Numbers) | Linked Standards (Line Numbers) |
|---|---|---|---|---|
| 1 | APIARY / SALIX | | | |
| 2 | Data Management Systems: SYMBIOTA, SPECIFY, BGbase, etc. | | | |
| 3 | SGR (Scatter Gather Reconcile) / FP (Filtered Push) | | | |
| 4 | GeoLocate; Geomancer | | | |
| 5 | ABBYY OCR (commercial); TESSERACT OCR; Adobe OCR | | | |

5. Define specific gaps that exist within each of the identified tools/systems (e.g., functionality problems, scalability limitations, availability, licensing issues, cost, lack of standard usage, missing features).

| Rank Order | Label Capture and Post-Processing Tools (list 1-5 from table above) | Gaps, Issues and Opportunities for Improvement |
|---|---|---|
| 1 | APIARY / SALIX | Licensing for OCR; May work better with proprietary OCR software; Accommodation of handwriting |
| 2 | Data Management Systems: SYMBIOTA, SPECIFY, BGbase, etc. | Gaps associated with the components |
| 3 | SGR (Scatter Gather Reconcile) / FP (Filtered Push) | Clustering algorithm improvement; Not ready for deployment (FP) |
| 4 | GeoLocate; Geomancer | Ability to store and reuse changes |
| 5 | ABBYY OCR (commercial); TESSERACT OCR; Adobe OCR | Handwriting; Licensing costs |

6. Identify the critical implementation date for HUB appliances that would enable/enhance label capture and post-processing based upon TCN project plans. Explain why this date is critical.

| Critical Implementation Date (Appliance) | Explanation |
|---|---|
| TSTB--TTD: 7/1/12 | Subcontracts for entomology begin on this date (7/1/12). For botany, their contracts begin 1/1/13. |
| LBCC: ASAP | Work is already underway; if it's not available from HUB, we will have to complete the work ourselves. |
| IN: 1/1/13 | Efforts now are concerned with imaging/raw data capture. |

---

7. Produce documentation related to the development/implementation of a label capture and post-processing appliance to serve the needs of the ADBC community.

| | |
|---|---|
| **Functional Requirements:** | Image processing, image management (local storage/management and transfer to HUB) and text processing.<br>Modular components (for some) vs. total "do-it-all" package (for others); |
| **Estimated computational resource requirements (computation, storage, network capacity):** | Storage space -- are label images stored here?<br>TBD. Requires additional knowledge of how appliance will be configured/deployed. |
| **Specific items the HUB needs to deliver to enable/enhance label capture and post-processing:** | Post-processing appliances; packages/workflows -- what will be hosted at TCN vs. what will reside at HUB<br>Image data management mechanism<br>Feedback/tagging mechanism (for OCR)<br>Useful tool: image "segmentation" software<br>Quality assurance/control module |
| **Specific items the TCNs needs to deliver to enable/enhance label capture and post-processing:** | Unique to a TCN:<br>TTD: lightbox -- share best practices<br><br>Applicable to more than 1 TCN:<br>Best practices for imaging activities |

Provide a risk assessment related to this label capture and post-processing appliance.
***Likelihood of Occurrence: 1 = Highly Likely, 2 = Somewhat Likely, 3 = Not Likely***
***Impact of Occurrence: 1 = Significant Impact, 2 = Moderate Impact, 3 = Little/No Impact***

| **Risk Name** | **Brief Description** | **Likelihood of Occurrence** | **Impact of Occurrence** | **Potential Mitigation Strategies** |
|---|---|---|---|---|
| Data error | Incorrect data capture by OCR | 1 | 1-3 | controlled vocabularies; post-processing curation (e.g., crowd-sourcing, etc.) |
| Destructive imaging | From handling | 2 | 1-3 | training modules; oversight; best practices |
| Data loss | e.g., corruption during file processing/transfer | 2-3 | 1-3 | "versioning"; backup (redundant storage) |
| Data integrity | Lack of adherence to standards or best practices;<br>Verbatim vs. Interpretation metadata tags | 1 | 1-3 | Best practices, standards, versioning |
| Software sustainability | Software provider goes out of business or discontinues software | 1-3 | 1-3 | |

8. Other notes, comments and details not captured elsewhere.