Enabling the TCNs and Collaborators
Breakout Group #2: Data Management & Archival

Facilitator Name: Brian Wiegmann

Scribe Name: Maribeth Latvis

Time Allotted: 150 minutes

Group Participant List: Thomas Nash III, Nahil Sobh, Linda Gruber, Katja Seltman, Elizabeth Martin, Jim Beach, Marcia Mardis, Alex Thompson, Jose Fortes, Kate Rachwal

Objectives:
Discuss and produce a report to summarize specimen collection data management and archival needs 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 data management and archival of specimen data (#1 is the most critical challenge).

Rank Order	Challenges Related to Data Management and Archival of Specimen Data
1	**Define roles**. Our goal is to enable integration, but not to curate the data. We need to define these roles between TCN/HUB. What are the TCNs not doing that could be helpful down the road - how do you know what they are *not* doing?
2	**GUID persistence and tracking**. Unique identifiers. The community needs to buy into it.
3	**Data backups**/ redundancy (action item Nahil will lead)
4	**Best practice guidance**. Data standardization (beyond Darwin core).
5	**Data quality**
6	**Storage location**. iDigBio in short term, but we need a long term plan. Data curation and authority: The need for "virtual curators" for these virtual databases - data longevity.
7	**Accessibility**: how easy and quick is it to access the data (generally and within projects)? Many sources of data are out there that are very useful, but aren't yet accessible.

Others (non-prioritized list)	Storage of raw data vs. transformed data? Data management in process. Technical training A bidirectional interface back to TCN. Synchronization of updates of data/annotations. Feedback pathways through the portal. Maintenance and leveraging authority files Tracking specimens through the network (keeping identifiers consistent -> education) Audience identification Data size prediction (image files, etc.) Data logistics General software support Object versioning- archival Existing databases (ITIS, Encyclopedia of Life) do not have infrastructure in place for **efficient updates**. Modular perspectives to find solutions: Ex. a module of geographic names that could be incorporated into a workflow could be helpful to other groups who don’t already have it. **AUTHORITY FILES: (action item, Katja will lead)** **-Communication between TCNs** to discuss shared problems (eg. with authority files). Are these issues being documented for posterity (blog format or wiki are options)? Moving forward into working groups. **-Need to integrate databases:** Consistent authority files (taxon specific databases) across different projects. Currently, different workflows exist for different projects (eg. plants vs parasitoids). How to merge these data down the road? Most people are without existing databasing systems, so will be flexible and open to efficient solutions.

2. Identify and order up to five existing practices and techniques that can be leveraged for data management and archival (#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	**Data Management and Archival Practices and Techniques**
1	**Barcoding (and other standards: ISGN geological specimen tracking)**
2	**Use of authority files (in use)- Expert validation**
3	**Mapping for data integrity (incl. georeferencing)**
Others (non-prioritized list)	**De-duplication (purging duplicates)** **Distributed object storage** **Outlier identification (existing quality control checks)** **Image search** **Collection ontologies** **Phenotype statements on specimens** **Exporting data to GBIF or using DIGR** **support of non-English URIs**

---

3. Identify and order up to five existing standards that can be leveraged for data management and archival. 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	**Data Management/Archival Standards**	**Explanation of Selections**
Other (non-prioritized)	(several listed, not ranked) Darwin core Audobon Core Apple Core OAIPMH XML EML FGDC Image standards (eg. jpg) NEXUS web service standards (JSON)	

---

4. Identify and order up to five existing tools/systems that can be leveraged for data management and archival (#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	**Data Management and Archival Tools**	**Explanation of Selections**	**Linked Practices/ Techniques (Line Numbers)**	**Linked Standards (Line Numbers)**
Other (non-prioritized)	(several listed, not ranked) Filtered push Specify Google Refine Open Stack Swift Geolocate Morphbank Symbiota Salix Medici			

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	Data Management and Archival Tools (list 1-5 from table above)	Gaps, Issues and Opportunities for Improvement
1		GUID architechture - Authority file updates
2	**Measures of data quality:**	Do we have a way of validating our product? Some files will be more uncertain than others (*Genus c. f. species*), and we should not ignore this uncertainty. Was the label legible? Darwin core has a comment field for each record for this information, although there is no standard.
3		Messaging infrastructure
4		Helpdesk/Learning - web service for data entry?
5		others, unranked: APIs Software development/ Hackathons (the HUB has this role?) International georeferencing OCR, Handwriting analysis software crowd sourcing tool Species file (for authority files)

---

6. Identify the critical implementation date for HUB appliances that would enable/enhance data management and archival based upon TCN project plans. Explain why this date is critical.

Critical Implementation Date (Appliance)	Explanation
Now Now April 2012 June 2012 now- June 2012	GUIDs communication about building authority files each TCN should send a preliminary set of digitized data. This would force the emergence of a mechanism to share data. storage and backup decision tool delivery -> timeline of specifics will require further discussion

7. Identify the critical implementation date for agreement to common data management and archiving standards between the HUB and TCNs/Collaborators. Explain why this date is critical.

Critical Implementation Date (Standards Agreement)	**Explanation**
Now	Decisions about authority files -> now (identifying what should be an authority source, collaboratively edited?)

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

*****OTHER ACTION ITEMS**

A facility where the TCNs can upload small test datasets:
-specify interfaces and standards that the TCNs converge on (database examples).
-provide a filter so that they have the same structure.
    -the iDigBio HUB can serve in harvesting and vetting existing databases.