Enabling the National Resource
Consolidated Results from Three Separate Groups
(Note that priority values may be duplicated due to consolidation and tied-rankings within groups)

Facilitators: Alan Prather, James Macklin, Jim Beach

Scribes: Grant Godden, Jill Holiday, Maribeth Latvis

Time Allotted: 195 minutes (each group)

Consolidated Participant List: Corinna Gries, Nahil Sobh, Linda Gruber, Melissa Tulig, Katja Seltman, Elizabeth Martin, James Hanken, Nelson Rios, Lucinda McDade, Greg Riccardi, Alex Thompson, Bruce MacFadden, Kate Rachwal, Edward Gilbert, Thomas Nash III, Christopher Dietrich, Christine Johnson, Robert Naczi, Austin Mast, Deb Paul, Gil Nelson, Kevin Love, Andréa Matsunaga, Jose Fortes, Shari Ellis, Matthew Collins, Zack Murrell, Barbara Thiers, Umberto Ravaioli, Jeffrey Holland, Randall Toby Schuh, David Bloom, Brian Wiegmann, Marcia Mardis, Pam Soltis, Renato Figueiredo, Reed Beaman, Betty Dunckel

Objectives:
Discuss and produce a report to summarize key issues related to enabling the National Resource. Nominate a reporter to deliver a 15-minute summary report to the plenary session at the conclusion of your session.

Deliverables:

1. Identify as many Stakeholder organizations as possible in fifteen minutes. Identify one key individual within each organization when possible. Consider a broad range of domains, including but not limited to the biological research community, standards/tools organizations, local/state/federal government agencies, non-governmental organizations, educators, students, and the general public.

Stakeholder Organization	Key Individual(s) Within the Organization
AAM (Association of Museums)	
Academia at all levels (college university and K-12)	
AFS (Field Stations)	
ALA	
All biological societies	
ASTC	
BIEN (part of iPlant)	Brian Enquist
Biodiversity Researchers	
Biology Database	Shannon Peters
BLM	
Bureau of Land Management (Interior)	

Bureau of Reclamation (Interior)	
CBOL	
CCH/SEINet -- Large-scale data shareholders	Dick Moe; Les Landrum/Tim Lowry
COL	
Commercial Service Providers (eg. web consultants)	
Consortia: taxon-specific (have portals, concerned with data/best practices; can be used as regional organizing pools)	
CSA (Citizens Science Alliance)	
Discover Life	
DNR (State specific)	
DOD - Department of Defense	
DOE - Department of Energy	
DOI Climate Centers	Damien Shea; Douglas Beard
EoL	
EPA	
ExEP (Exotic Plants)	
GBIF	Donald Hobern
GNA	
GSCG (UN, CBD)	
Homeland Security	
IALE (International Assn Landscape Ecology)	
IAWGSC (inter agency working group on scientific collections)	
Informaticians	
Interagency working group on scientific collections	Scott Miller
iPlant	
ITIS	Stinger Guala
IUBIO	
IUCN	
Land Grant	
Lifewatch/ Framework 7	
Local level: Parks, Cities, LEAs (local education agencies)	
LTER	
MorphBank Standards	Greg
Museum Curators	
NABT (National Assn. Bio Teachers)	
Nat. Phenology Network	

National Park Service (NPS)	Anne Hitchcock
Nature Serve	
NEON	
NEOTOMA	Russ Graham
NGOs - Nature Conservancy (and other conservation organizations), Nature-Serve, Botanical Gardens and Museums, Cultural Institutions, Audobon, World Wildlife Fund, Sierra Club, Defenders of Wildlife	
NIH- NCBI	
Non-TCN organizations (e.g. VertNet);	David Bloom
NSF	
Politicians, public officials, policy-makers, lobbying groups	
Private citizens, citizen scientists, local and regional societies, clubs and centers (Bug guide, Wikispecies)	
Professional Organizations: e.g., ABLS -- American Bryophyte and Lichenological Society	Committee
Professional Societies (BSA, ESA, AIBs et al)	
SCICOL	David Schindel
Species file	Matt Yoder
SPECIES LINK (“Brazilian GBIF”)	
SPNHC	Tim White
State - AID programs	
State Heritage Programs	
State level: Departments of Natural Resources; Conservation; Transportation; Wildlife; Natural Heritage; Division of Parks and Forests	
TDWG (taxonomic database working group)	Chuck Miller
Teachers and Educators - e.g., NSTA	
TNC (The Nature Conservancy)	
UNESCO	
United Nations	Edward Morton
University Administrators	
US Fish and Wildlife Service	
USDA ARS	
USDA-APHIS	Ann Bartuska
USDA-NRCS	
USDA-CSREES	
USFS (Forest Service)	
USGS	Stinger Guala, Kurt Ridder
USGS -- Core Science Analytics and Synthesis	Lucy Edwards

USUH	Zack Murrell
Volunteer based organizations - (e.g., AARP, Americorps)	
WHO	
Zooniverse (Crowdsourcing)	
Zoos	

---

2. One focus of iDigBio is to produce a portal that integrates collections data from many different institutions, and to make that data searchable/accessible via a website. In a brainstorming session, define requirements for the "iDigBio Version Zero" platform (the first, rapid iteration of iDigBio integration and search functionality). What limited standard(s) and common data format(s) should be supported to enable iDigBio to access data from multiple collections through a single portal, with a single query, with persistent resolvable globally unique identifiers, with results appearing as one, within six months?

Selecting existing digitized specimen information to provide advanced capability in advance of TCN digitization. What are important databases that we might consider? Examine characteristics of the data (from each) that should be considered.

What is the target audience?

Series of feedback mechanisms and commenting capabilities.

Visualization capabilities for data that will enable quality control for TCNs (e.g., georeferencing, taxonomy). Couple this with feedback mechanisms/capabilities.

Help desk functions.

Clearinghouse of relevant information and community activity.

Ecological approaches/routes into data. Novel approaches to data exploration.

Portal for all data gathered from TCNs.

Ability to facilitate data conversion by automation.

Data movement/transfer to/from TCN and HUB.

Attribution and copyright for images (metadata).

Caution: This is assuming that we will have data - GBIF's cache, net's and iss's and consortia and other people that have an automated interface

"What will it look like"? 1. Query box (simple and advanced). Types of queries? Terms associated with specimen: taxonomy, locality, substrate, collector name

Ability to export file like GBIF (simple queries and download of data for people to take away)

Usage pattern tracking

Possible web service, coordinate => species list

Samples of searching: collection date, collector (ok), taxon, location, institution/collection

A map. An image thumbnail & full size. Search by point radius, locality in addition to specimens.

A map of all things near a point ala zipcodezoo - syndication and acceptance of observations

Link scientific name to EoL, perhaps link other things to other sites. Links to TCN's sites w/ further information such as digital publications.

Catalog of life - has an API for taxonomic name synonyms (EoL too maybe) -> Probably our search in 6 month will be just string matching

Need documentation for iDigBio APIs to put in front of potential tool builders

Can I log in with someone with roles? Can we reason about what people would be interested in and push w/ RSS?

RSS -> feedback about who downloaded data

Aggregated page

Provenance to reuse patterned workflow/processes

6 month bottom line: collect "easy" data, make it searchable, attach an image, add to map

The least common denominator is Darwin Core, which GBIF uses as an aggregator. This can be used as a starting point. Do we need a standard that doesn't exist yet? It would be a shame to lose historical data (i.e. from those specimens without coordinates).

Is it reasonable to have a portal out there in 6 mos or should we focus on internal work flow improvements?

Specimen Standards
- How do new TCNs interact with the HUB?

- Intellectual effort -> best practices doc, data formats.
- Validation and “enforcement” effort

Differentiation from GBIF
- Why are we reinventing GBIF? GBIF just searches and maps... we should provide more.

Proof of Concept or prototype web portal (e.g. “bees of Canada”, perhaps using barcoding project as template)
- Can be “quick and dirty” … yet searchable and improvable
- Proof of scalablity (in terms of performance, capacity)
- e.g. symbiota and lichen group
- focus on some taxon, “named collection” or geographic aggregation (e.g. biota of Florida, a project that is already in progress at FLMNH)

Compromise of internal and external value-add services

Is this a publishing front for this project or a TCN database? Should the focus be inward or outward?

> EXTERNAL: public search. Serve as advertisement of our initiatives. (in 1 year?). Stakeholder meetings should take place and their needs should be incorporated.
>
> INTERNAL: data downloads through web services (at the end of 4 years).
> Portal should exist as soon as TCNs are generating data.
> Standards
> Pipelines
> Services for data uptake

---

3. Identify at least five communication challenges between the HUB and TCNs/Collaborators. For each identified communication challenge, present at least one recommended strategy or technique for overcoming that challenge. For example, BIO stakeholders and IT stakeholders often speak different languages. Can this be overcome through subcommittees that act as liaisons for communication between the groups? What other techniques would be effective?

Rank Order	Communication Challenge	Recommended Strategies or Techniques for Overcoming the Challenge
1	real time communication between HUB/TCN	web-ex, wiki, video-conferencing, desktop sharing (BigBlueButton, GoToMeeting, Skype)
1	Clear communication of projects and timelines (benchmarks)	Process maps (including project tools and lists of milestones); Project tracking tools for TCNs via the HUB
2	biologist/IT communication (tool/resource specific) OR where (to whom) to direct a question	FAQ by biologists/FAQ by IT clearinghouse with answers (archived questions)/list-serv/blog
2	User needs assessment; Use case development; Opportunities for input for tool building	Open source development techniques -- public bug detection; feature requests Development teams that include biologists
3	virtual meetings are not always feasible	IT folks need to be able to communicate regularly (formally or informally)
3	Identification of contacts and breakdown of responsibilities	Defined working groups with an identified contact; Point person at each TCN and a point person at iDigBio that is assigned to each TCN;
4	Sustained communication flow	Monthly skype meetings or web conferences (e.g., 1-hr meetings where people can communicate concerns/ideas)
Unranked	Acronym proliferation	Wiki Glossary/Dictionary
Unranked	Avoid perception of top-down	Inclusive decision making via working groups
Unranked	Communication between TCNs	Workshops, working groups, etc.; Use of communication structure at iDigBio -- e.g., publication of best practices

Unranked	Other stakeholders don't know who to contact	iDigBio liasons to stakeholders
Unranked	Acronyms as an impediment to communication	common language/terminology/vocabulary -glossaries (for training, techniques, general understanding) -avoid acronyms (or provide "dictionary" of acronyms for clarification)
Unranked	Define role of the HUB: coordinating communication to broader community. -> TCNs are widely distributed. -> Disparate levels of communication. -> proactive editing and moderating of communication channels **Resource limitation hindering workshops (can isolate those groups that do not have funded projects, yet are pursuing similar objectives) Not all groups are funded through the same sources, creating conflict. Only enough money to provide a backbone.	A tree structure for where information needs to go. Modes for inclusive communication (everyone needs to feel included. Widespread ownership and engagement). Many different levels needing support (trench workers, PIs, prospective TCNs, interested 3rd parties, etc... each needing different modes of communication) -> newsletter (NESCENT has one, potential template. Good for PIs) -> forums, blogs, wiki, listserves (better for frequent users) -> social media to post new ideas (Twitter, Facebook). A "post book" model to empower involvement at lower levels (undergrads should feel comfortable presenting new ideas. -> Needs seeding and fresh content -> Should facilitate collaborative work and provide tips, Q and As **a survey to assess issues and possible solutions?** fold all groups into the strategic plan. Defining a community name (ADBC).

Unranked	Communication from TCNs back to HUB	Every 3 months reports are sent to HUB. These should be used to parse info, priorities. Mechanisms besides reports (reports are thought of as progress reports): workshops, wiki to continue conversations internal advisory committee
Unranked	working groups	forum (not currently active)

---

4. Identify and rank functional requirements for collections cross-referencing within the iDigBio portal.

Rank Order	**Functional Requirements for Collections Cross-Referencing within the iDigBio Portal**
1	GIS - georeferencing of specimens into layers for intersecting w/ other datasets
1	Linked data; Showing possibilities for linking
2	Matching of GUIDs
2	Ontologies of known relationships to infer other relationships
3	Data standards (e.g., geographic identifiers and interfaces; authority files)
3	Data mining algorithms to find patterns
4	Development of controlled vocabularies for shared data fields
4	Clustering/association of queries to infer relationships
5	Reconciliation of different authority files -- including spellings and abbreviations
6	Flexibility of authority file systems (dictionary? hierarchical or advanced search?)
7	Text matching functionality

---

5. Identify and rank at least fifteen digitization tracking metrics that should be provided by data contributors (e.g., source of funding for each digitized specimen, camera operator, publication, citation).

Rank Order	Digitization Tracking Metrics Provided by Data Contributors
1	Sources of funding
1	Tracking ownership / Credit (e.g., tracking methods for all contributors of data associated with each specimen)
1	Collection origin
1	price / specimen
2	Camera operator (photographer)
2	Publication/Citation
2	Ownership issues -- who entered data
2	Digitization method: 1) Hardware (image capture and post processing methods)
2	Digitization method: 2) Software (image capture and post processing methods)
2	specimens / time period / worker for productivity tracking
3	Versioning
3	Quality control / data verification fields
3	image collection & manipulation metadata
4	log files from software for errors, unexpected behavior
5	number of duplicates / specimen & compilation records
6	quality of image vs. quality of specimen (filtering if quality of image does not meet acceptable standards)
7	time to create quality image, harvesting duplicates etc. (efficiency/best use of time)
8	zooniverse and crowd-sourcing accuracy (e.g. Harvard Herbarium Home allows user to rank their confidence)
9	crowd - sourcing ranking statistics, “training” or how-to explanations from top-ranked users
10	who/when/what

6. Identify and rank at least fifteen tracking metrics that should be maintained by the HUB and made available to source data contributors (e.g., data integration statistics, update statistics, search result hit count, detail view count, query type count, count by portal – research vs public/educational).

Rank Order	Tracking Metrics Maintained by the HUB
1	Rate of museum-wide digitization capture (and/or individual productivity by subcontracts)
1	attribution statistics (who did it, rank of work (how many accessed)), institution of contributor
2	User statistics and search results -- e.g., hit counts (public data use vs. research use); Will we have known users? Will we allow downloading of data?; Tracking of query terms; Publication tracking (this will be made possible by providing citations for the data via the HUB)
2	type of research question/study (e.g. search terms or search parameters)

---

7. Define and/or illustrate an interaction model between the HUB and TCNs. Identify what functions/processes occur at the HUB, and what functions/processes occur at the TCNs and Collaborators. Identify points of interface between the groups.

What parts of the workflow will occur at the TCN vs. what will occur at the HUB?

Interactions: broader than communications

I. Existing TCNs:

Expect the HUB to communicate news, activities, plans, developments, status, etc. via regular updates.
Communicate resources and discovery
Work towards common goals re: sustainability
HUB -> TCN oversight responsibility (technical and productivity)

TCN (with their own objectives defined by interactions with clients) **--->** HUB
Quarterly reports: status/productivity

What incentivizes the two to communicate besides the reports? Working groups? Collaboration in seeking additional resources?
Do we have a "lax" model or a "strict" model (i.e. progress reports are posted)? Genome project model (with benchmarks) as a standard?

1) Report challenges and road blocks
2) TCN can expect and request technical support services

II. New TCNs, just funded:

Who is doing what?
Assume we are a new TCN that has just been funded (say, in year 4). What are the standards and protocols do we follow for our data to become acceptable to the HUB?
-> The TCN is expecting direction from the HUB for direction as to how to set up their project.

A **toolkit** of best practices with a help desk (differentiate "standards" from "best practices" and "techniques")
HUB provides the direction
Facilitate and engage with existing TCNs

Each TCN will nominate representatives to meet 1-2 times per month: an internal advisory board to fulfill such a role.

III. Retired TCNs: ("shadow TCNs" "hangers on")

"post-mortem" relationships with HUB, interested third parties
1) Continued curation

Most TCNs would probably continue generating data
Someone should still be in place (a curator) to mind these data

2) "Retirement Planning" or "Data Life Cycle Planning"
Data preservation
Expertise leveraging
3) Planning a "data legacy" upon data generation (eg. digitizing field books upon return from a trip for posterity) -> **promoting best practices**
Digitization should be proactive not retroactive

---

8. List at least five issues related to data/image rights for data elements indexed and published through the HUB. Offer a potential solution for at least 50% of your identified issues.

Rank Order	Data/Image Rights Issues	Potential Solution(s) for Data/Image Rights Issues
1	need best practices set up	see creative commons
2	fair use; literature	

---

9. Define key differentiators and value-add services that iDigBio should deliver. The following list is provided as a starting point for your discussion:

a) iDigBio will deliver data, images and media to the end-user.
b) iDigBio will provide long-term storage for images and data.
c) iDigBio will repurpose and deliver to all participants existing tools (e.g., analytics tools, imaging tools, database tools) that may be currently limited to certain collections portals. These tools will be packaged as "appliances" and delivered to the collections community.
d) iDigBio will provide full access to the entire record source directly within the iDigBio portal. There will be no need to point back to the source for extended data elements.
e) iDigBio will allow data providers to report and analyze end-users accessing their data.
f) Specimen records displayed through iDigBio will be tagged with source information for downstream ownership identification.
g) iDigBio will facilitate reporting back to data providers regarding problems with their data via annotations (e.g., incorrect species, incorrect geo-referencing). Feedback will become publicly available immediately and final data changes will be updated/corrected/rejected by the data provider.
h) iDigBio will utilize persistent resolvable globally unique identification for each record/object that will not change when source data is refreshed.
i) iDigBio will integrate specimen data and images only. Occurrence records not tied to a specimen will not be included.
j) iDigBio will provide programmatic interfaces to query data.
k) iDigBio will enforce data quality controls (e.g., required information, minimum imaging quality, automated cleansing).
l) iDigBio will archive integrated data.
m) iDigBio will automatically republish data to national/global aggregators.
n) iDigBio will provide data processing services.

Breakout Notes:
j: Expanded to include web services for quality control
l: Question: What is meant by "integrated data". Answer: "Integrated is the preferred term for the manner in which iDigBio's portal will collect and manage data, distinct from an "aggregator" that may not maintain persistent relationships.
m: Solve how to keep the latest version out instead of multiple versions?

Rank Order	Key Differentiators and Value-Added Services That iDigBio Should Deliver
1	all of the above
2	republishing is a problem (dupes) and attempting to create unique identifiers is still problematic
3	All coordinated activities should be done by the hub
4	We believe that the following services are the most important: h, a, b, k, j

---

10. In a brainstorming session, design an educational outreach website that would be provided on the iDigBio portal that utilizes specimen data and images.

Define your target audience (e.g., K-12, undergraduates, informal education consumers).

Informal science education consumers -- non-adults.

Define the educational objectives of your appliance.
Incorporation of two themes:
1. Introduction to local biodiversity
2. Introduction to collections (and historical components)

Define the features of your appliance.
Notes:
Mobile apps -- e.g., for local biodiversity discovery
Essay contests / Games
Videos
Building games
Linkage of relationships among biodiversity
Interactive
Hac-a-thon Experiments
Use cases to train non-systematics community in data usage potential.

Separate Recommendation: A crowdsourcing site, and/or K-12 (teaching modules/activities/georeferencing)

---

11. Identify and rank at least five critical topic-focused workshops that must follow this Summit. List at least five key personnel who should attend each identified workshop. For each identified topic, note if there is sufficient complexity and time required for resolution that it should be handled as a Working Group that will meet regularly, rather than as a single workshop.

Rank Order	Workshop Topic	Working Group Required? (Y/N)	Key Personnel
1	Imaging standards and protocols (streamlining processes)	Y	Bob Morris; Gregor Hagadorn; People who have done a lot of imaging (including international members -- e.g., global plants initiative)
1	Augmenting OCR (natural language /post-processing/label and specimen data); Imaging	Imaging (use experienced resources outside our community)	Read Beaman, Ed Gilbert, Jason Best, US Postal Service, Group from Germany (James Macklin knows who), CIA, banks, John Hart CS professor, Xerox, IBM, Google, BHL Chris Freelan, library community, LoC; Nathan Wilson mushroom server Can this be colocated with a broader conference?
2	Authority files (to discuss development of appliance that can deal with authority files)	Y	Bill Piel; Stinger Guala; John Wieczorak; Matt Yoder

2	standards and interoperability (e.g. Expanding core)		John W., James Macklin, GBIF-Dave Rempson, TDWG, Greg Riccardi, library community
3	Georeferencing	Y	Nelson Rios; John Wieczorak; Carol Spencer
3	Identifiers and persistence		Greg Ricardi, GBIF best practice papers have names, ISO representative
4	Data storage, curation, and movement	Y	Michelle Butler (NCSA); Chris Jordan; DataOne
4	crowd-sourcing/volunteerism/citizen science community building		Arfon Smith, EoL, WikiSpecies, Herberia@Home, Bruce McFadden, Michael Giddens, Vince Smith (vibrant), Cornell representative, Bugguide - John VanDyke, Earthwatch representative, ReCaptcha
5	Outreach and volunteer interaction (technology for crowd-sourcing) -- Across TCNs	Y	Tom Nash; Austin Mast; Michael Gibbons; Betty Dunckel; Nelson Citizen Science Alliance; Rick Bonney (citizenscience.org)

5	Sustainability		Dan Stanzione, Nescent, DataOne, xsede, DataNet, TerraGrid, all NSF funded
6-7	Workflow needs; General digitization working group		Jim Beach; Linda Ford (Rod Eastwood); Vince Smith
6-7	Education and Outreach		Carolyn Lewis; AMNH; Philadelphia Museum; Gaye-Lynn Clyde Milwaukee Public Museum; Joe Cook; Carolyn Ferguson; Anna Monfils; one rep from each TCN.
Unranked	Data mining (e.g., use of data for application to research questions)		
Unranked	Mobile access to databasing		
Unranked	Data and Image Rights		
Unranked	Education directorate within NSF Outreach: education specifically; developing biodiversity/informatics (short course?) workflows (imaging, packaging, file transfer etc) Engineers to help think about problems virtual workshops!!! communication funding and development	taxonomy cyber-infrastructure	
Unranked	Usability (not us)		
Unranked	Systematists/Biologists		
Unranked	Training of IT/Biologists for future generations		
Unranked	Biodiversity informatics		

Unranked	Specimen/Collections Data Standards -activities and status -reexamination of vitality and relevance -wrap up loose ends and finish by certain date	Y	
Unranked	Image and Media Standards -relevance, tools (SPNCH)		
Unranked	Collaborative Development		
Unranked	Technological “seamstress” -stitching together gaps -workflow task optimization		
Unranked	Stakeholder meetings		
Unranked	Website Design		
Unranked	Education and Outreach Options		
Unranked	Label Image Acquisition and Post Processing		
Unranked	Paleo participation		

---

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

- Educators are needed to address #10 above.
- Data rights must be discussed soon, since it will be a critical issue.
- Development of a policy for masking sensitive data -- propose policy and facilitate feedback via some means (non-working group).