

iDigBio TCN IT Needs Assessment Results

José Fortes,
Shari Ellis,
Andréa Matsunaga

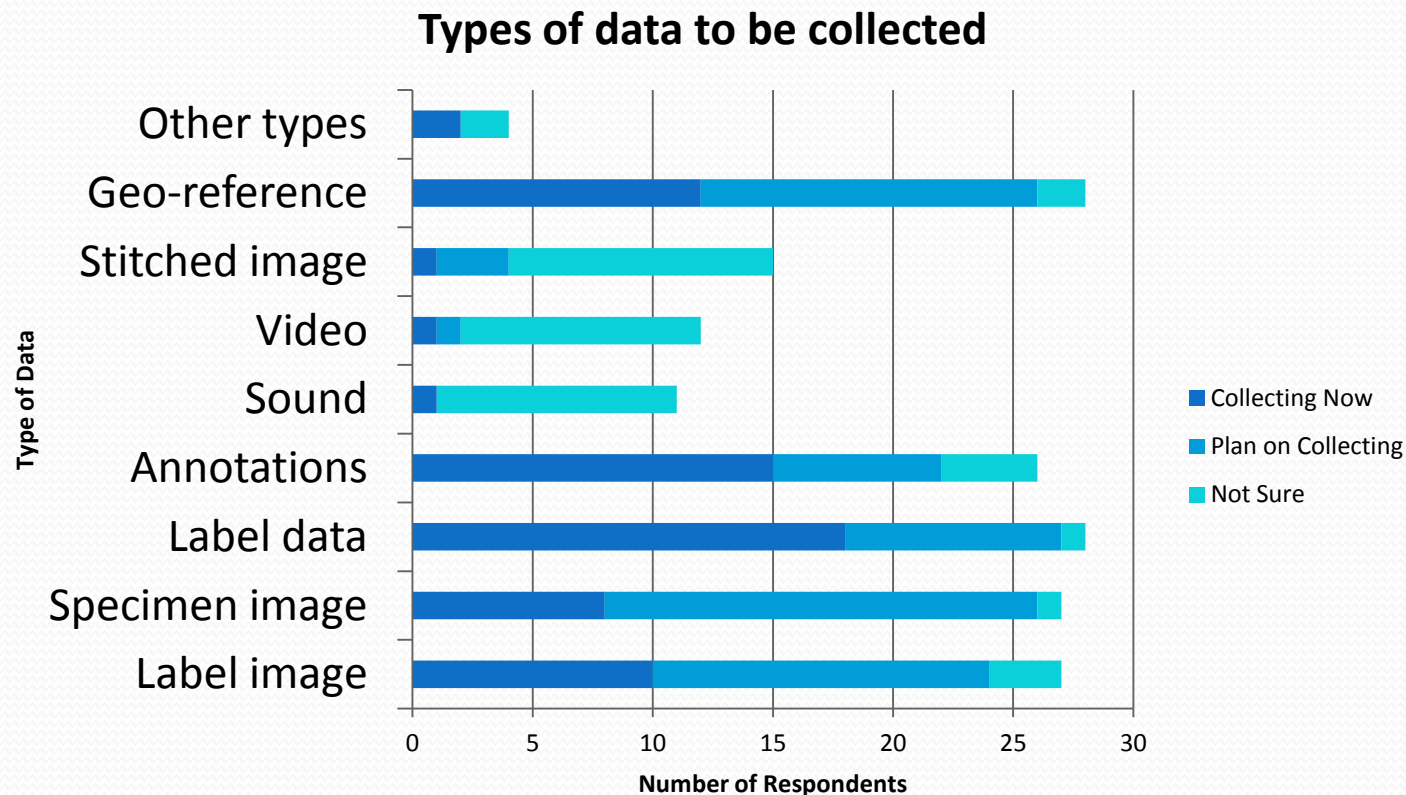


Overall Statistics

- 15-question survey of institutions in 3 funded TCNs
- 30 returned surveys (13+8+9)
- Significant variety in the types and amount of data being collected, storage requirements, and the software tools and services being utilized
- Three areas where the TCNs may readily look to iDigBio for guidance and support:
 - Archival storage
 - Feedback on the data
 - Computationally intensive transformations

Types of Data Collected

- Most collected: Label image and data, specimen images, geo-reference, and annotations

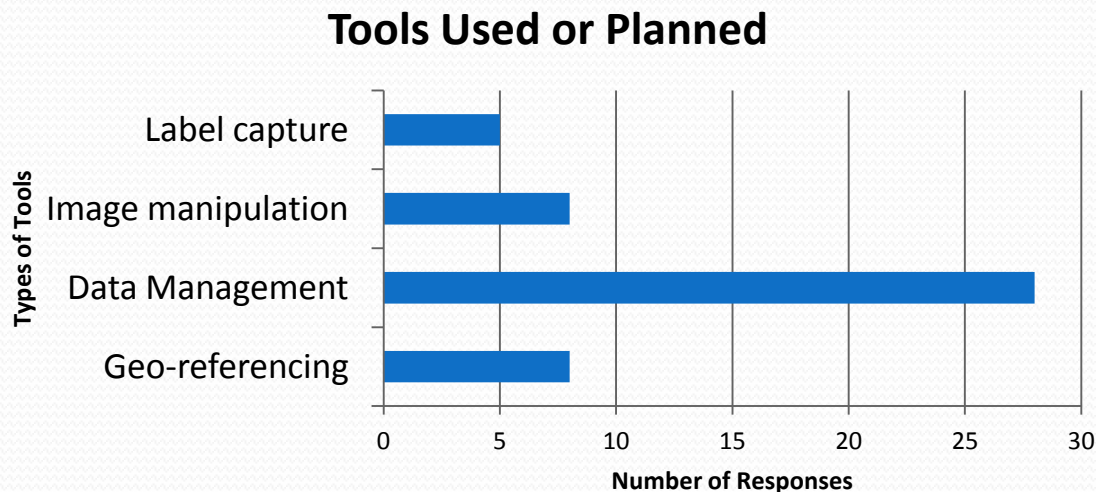


Metadata Collected

- Most often expected to be collected:
 - Operator or data enterer
 - GUID (globally persistent unique identifier)
- Other metadata included:
 - Source of funding
 - Instrument
 - Light source
 - Copyright
 - Date of entry
 - Institution
 - Collection permits
 - Accession data
- None (2)
- TBD (4)

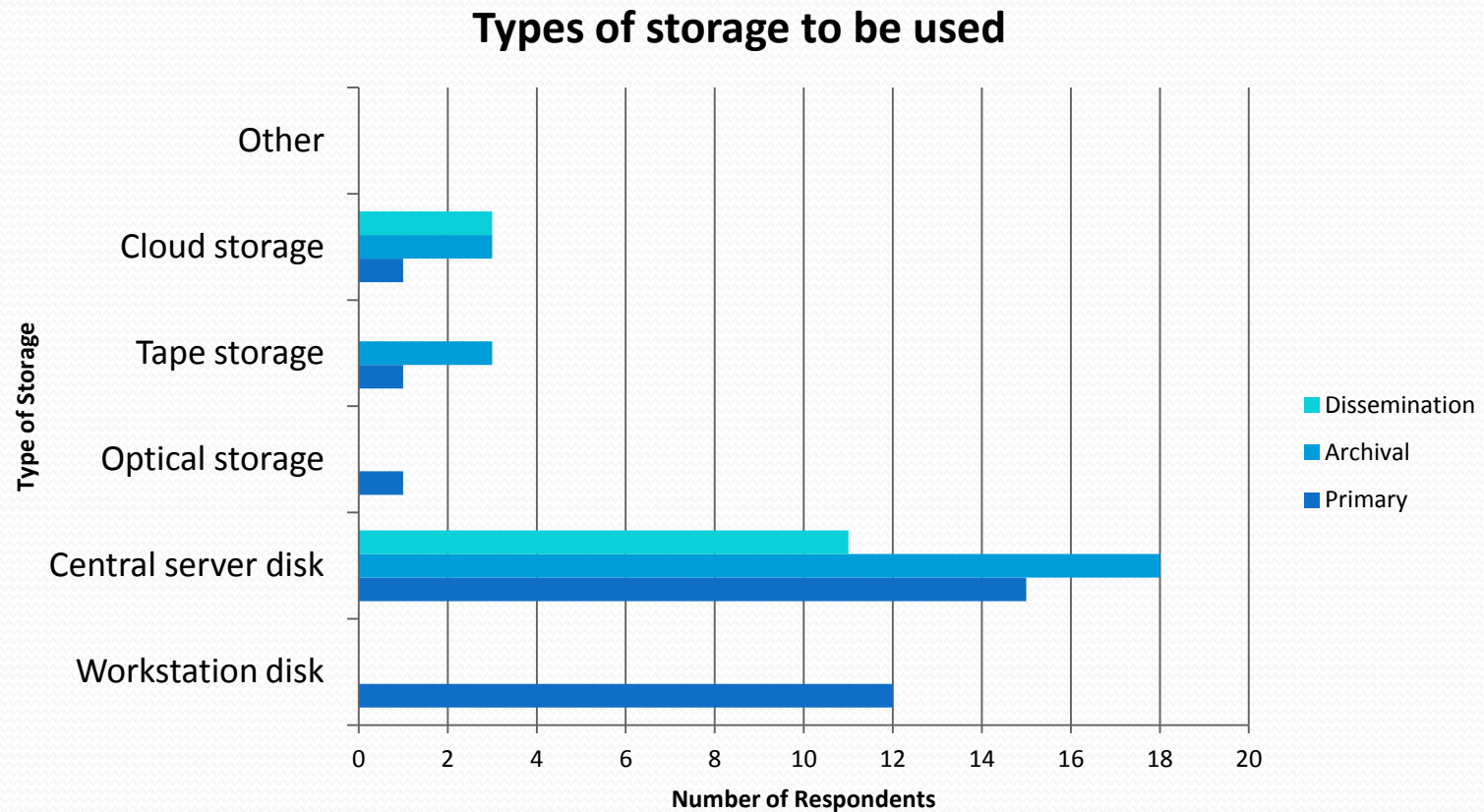
Tools Used or Planned by TCNs

- Label Capture:
 - OCR, OCR+NLP, Tesseract, SALIX, HERBIS
- Image manipulation:
 - Capture One (2), ImageMagick, Gigapan (2), DSLR camera software (2), Epson scanner software
- Data Management:
 - MS Excel (4), MySQL (1), MS SQL (1), FileMaker (2), MS Access (4), KE Emu (2), Accumulus (1), Silver Biology (1), Specify (9), Arctos/Oracle (1), Symbiota (2)
- Geo-referencing :
 - BioGeomancer (4), GEOLocate (3), Google Earth



Disk Storage Types

- Most common for primary storage, archival storage and dissemination: Central server disks



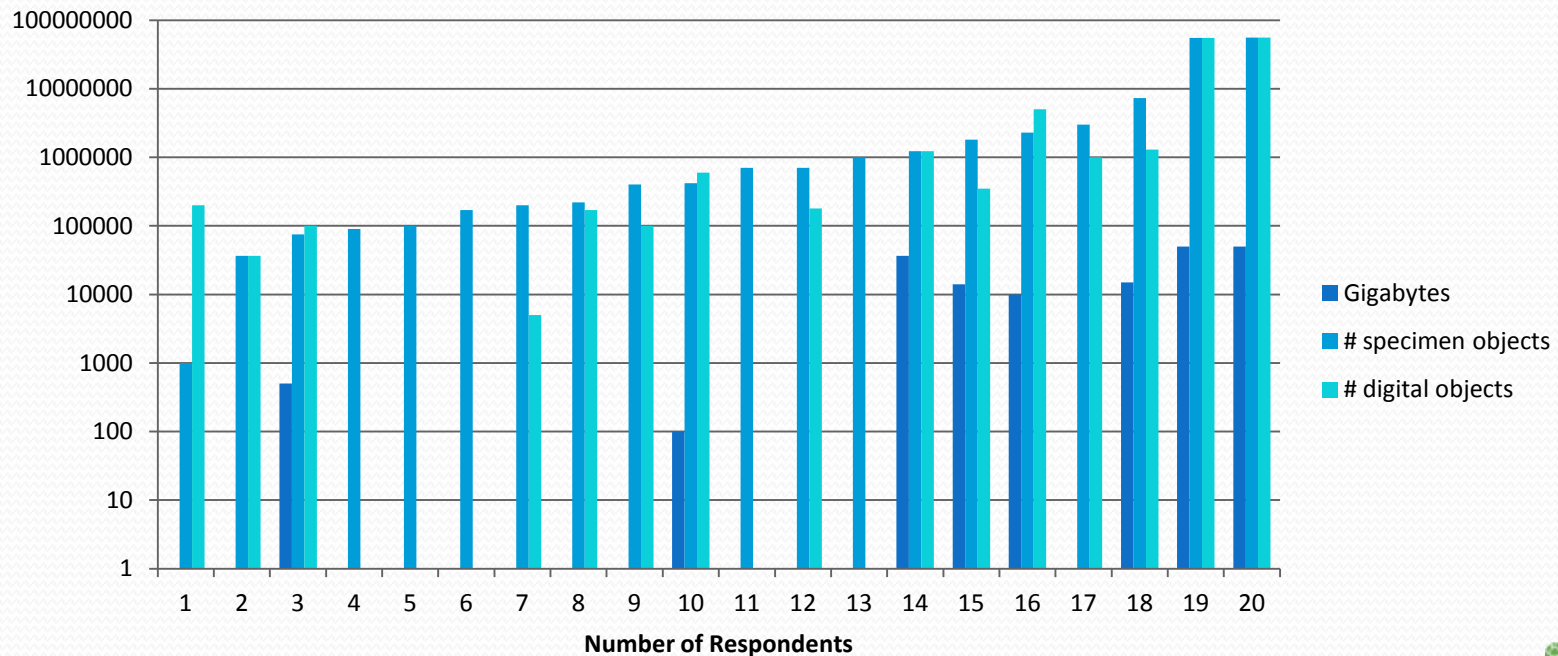
Tools for Primary and Archival Data

- Primary:
 - Data Management + user interface (desktop or web)
 - Arctos, Dspace, Fedora Commons , KE Emu, Silver Biology, Silver Collections, Symbiota
 - Data Management:
 - FileMaker
 - Data processing:
 - Medici
 - Raw data
 - NAS, FreeNAS, Delta Cloud API (AWS/Rackspace/NCSA)
- Archival:
 - CrashPlan
 - Backup Exec
 - Institutional SAN
 - Off-site server and own scripts
 - Expecting recommendations
- Unknown
- To be determined

Storage Requirements

- Large variation:
 - Raw storage: 100GB ~ 50TB
 - Number of specimens: 1,000 ~ 56 million
 - Number of digital objects: 5,000 ~ 56 million
 - Digital object per specimen: 0.025 ~ 200
- Sufficient storage: 13 “yes,” 10 “maybe,” and 3 “no”

Storage Requirements



Replication Agreements

Provider	Tools, Services, Comments
DataONE	Why?
MorphBank	agreed to take all images, than problems arose with too many images too fast
TACC	agreed to take all images, than problems arose with too many images too fast
GBIF	DiGIR (2 respondents) DwC-A Files KE Emu
EOL	Data (3 respondents) Maybe Species descriptions
Other Providers	DiGIR SYMBIOTA Discoverlife.org OAI iDigBio

Protocol and Standards

Protocol	Number of Respondents
DiGIR	9
TAPIR	3
ODBC	3
JSON	2
Other	Symbiota My University services Silver Image XML-RPC XML

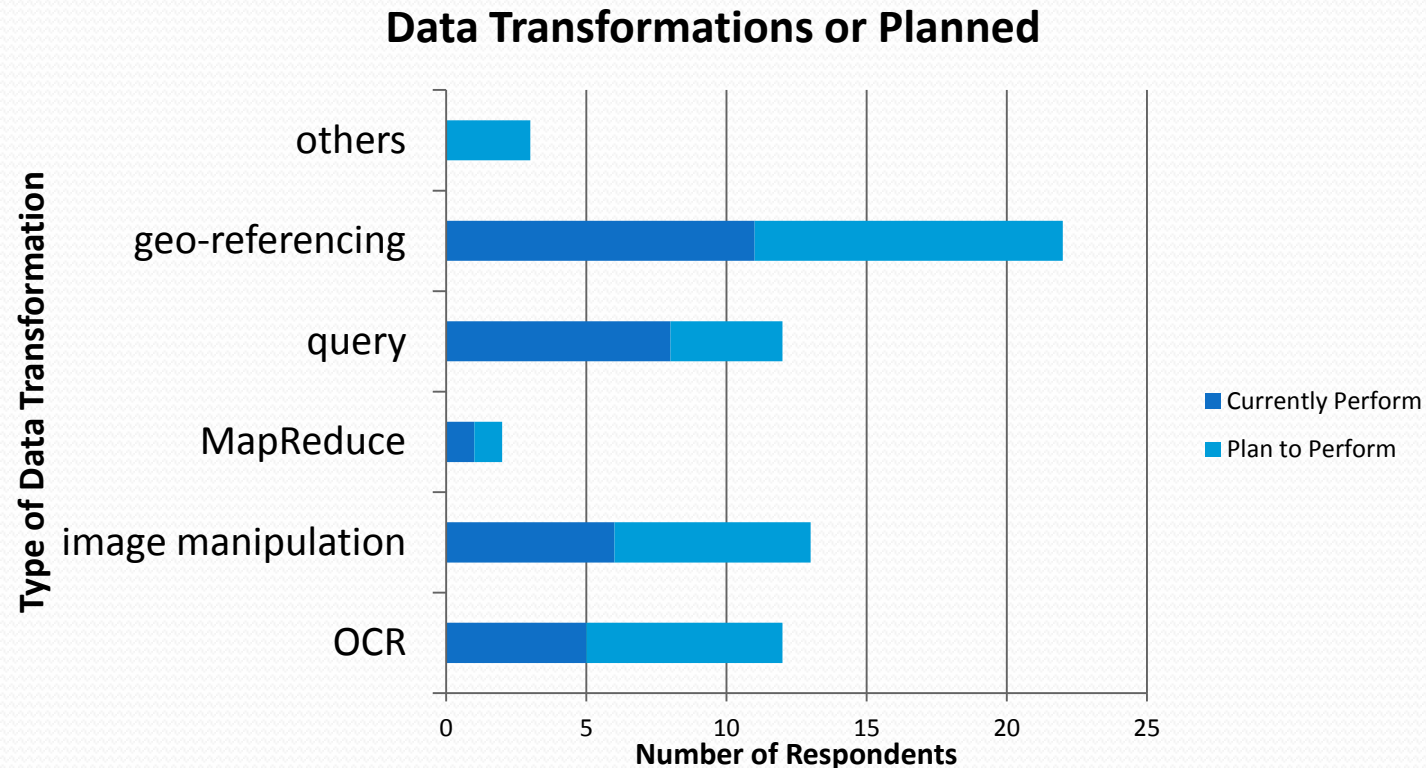
Standards	Number of Respondents
SQL	9
XML	8
DwC	5
KML	2
RDF	2
Other	My university standards Darwin Core Archive and custom text

Data Access

- Desktop:
 - SQL clients (3), Specify (3), own
- Web-based:
 - Discover Life, GBIF (2), HUBZero, Silver Collection (2), Symbiota (3), own (8), TBD (2)
- Feedback mechanism:
 - Email or comment forms (4), iDigBio recommendation (2), own suite of tools (1), TBD (10)
- Access Control:
 - No restrictions (6); sensitive locality data blocked (6); role-based access (5) and TBD (4)

Data Processing

- Most cited: geo-referencing
- Others: GIS analyses, imaging a subset of the collections, citizen scientists, custom data cleaning, and natural language processing
- From:
 - In-house (7), TCN collaboration (5), iDigBio (3), not an issue (2), TBD (4)



Questions?

Thank you!