IJDC | *Conference Pre-print*

Out of the Jar into the World! A Case Study on Storing and Sharing Vertebrate Data

Susan Borda
University of Michigan

Abstract

In 2018, the Deep Blue Repositories and Research Data Services (DBRRDS) team at the University of Michigan Library began working with the University of Michigan Museum of Zoology (UMMZ) to provide a persistent and sustainable (i.e., non-grant funded, institutionally supported) solution for their part of the National Science Foundation's (NSF) openVertebrate (oVert) initiative. The objective of oVert is to the digitize scientific collections of thousands of vertebrate specimens stored in jars on museum shelves and make the data freely accessible to researchers, students, classrooms, and the general public anywhere in the world. The University of Michigan (U-M) is one of five scanning centers working on oVert and will contribute scans of more than 3,500 specimens from the UMMZ collections (Erickson 2017).

In addition to ingesting scans, the project involved developing methods to work around several significant system constraints: Deep Blue Data's file structure (flat files only, no folders) and the closed use of Specify, UMMZ's specimen database, for specimen metadata. DBRRDS had to create a completely new workflow for handling batch deposits at regular intervals, develop scripts to reorganize the data (according to a third-party data model) and augment the metadata using a third-party resource, Global Biodiversity Information Facility (GBIF).

This paper will describe the following aspects of the UMMZ CT Scanning Project partnership in greater detail: data generation, metadata requirements, workflows, code development, lessons learned, and next steps.

Submitted 16 December 2019 ~ Accepted 19 February 2020

Correspondence should be addressed to Susan Borda, Shapiro Library 919 South University Avenue Ann Arbor, MI 48109-1185. Email: sborda@umich.edu

This paper was presented at International Digital Curation Conference IDCC20, Dublin, 17-19 February 2020

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

In June 2018, the Deep Blue Repositories and Research Data Services (DBRRDS) team at the University of Michigan Library began working on a project with the University of Michigan Museum of Zoology (UMMZ) to provide a persistent and sustainable (i.e., non-grant funded, institutionally supported) solution for UMMZ's part of the National Science Foundation's (NSF) openVertebrate (oVert)[1] initiative. The objective of this initiative is to digitize scientific collections of thousands of vertebrate specimens stored in jars on museum shelves and make the data freely accessible to researchers, students, classrooms, and the general public anywhere in the world. UMMZ is one of five scanning centers contributing to the oVert initiative and is slated to contribute more than 3500 specimen scans from its collections (Erickson, 2017).

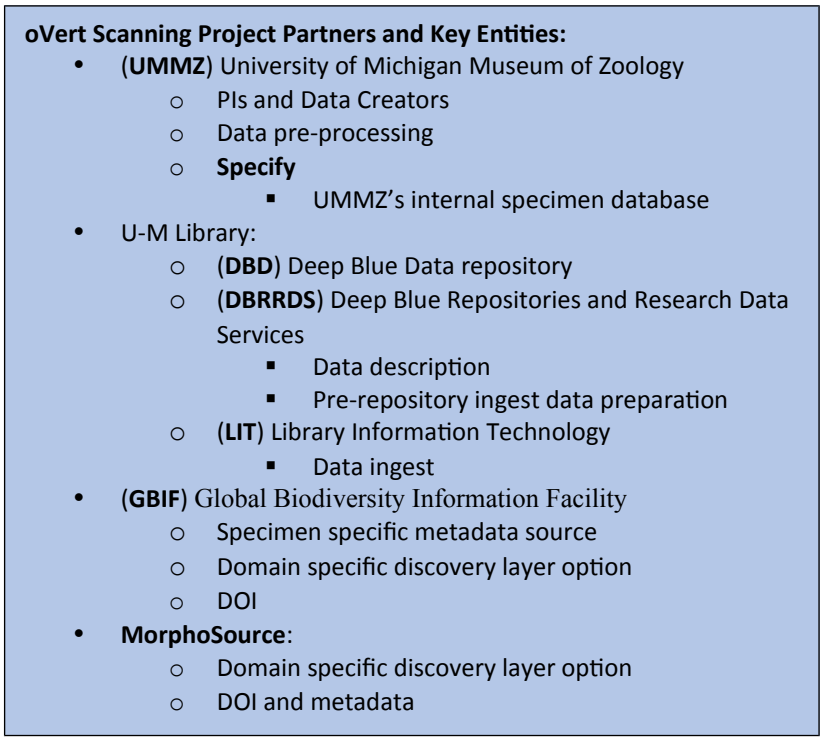


Figure1.Project phases

The project with UMMZ encompassed several phases, summarised in Figure 1. In the first phase, specimens were scanned with x-ray computed tomography (CT). In the second phase, the digitized images were ingested into an institutional repository managed by DBRRDS. Named Deep Blue Data (DBD), the repository is used to share and preserve data sets generated by researchers affiliated with the University of Michigan. DBD, is built on a Samvera Hyrax 2 platform. It can mint DOIs and handle large data sets (>1 terabyte).

The final, future phase will entail sharing the metadata from DBD with MorphoSource[2], a data archive for 3-D morphological datasets, for domain discovery.

Completing the second phase presented several significant challenges for the library. For one, the specimen scans could not be directly ingested to DBD in the structure created by the scanning process due to its lack of support for folder hierarchy. Therefore, the specimen data

[1] NSF oVert: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1701713
[2] MorphoSource: https://www.morphosource.org/

generated by UMMZ had to be represented as flat files. In addition, the UMMZ specimen database, Specify[3], is a closed database for specimen metadata with information inaccessible to non-UMMZ group members. To overcome these challenges DBRRDS had to create a completely new workflow for handling batch deposits at regular intervals, develop scripts to reorganize (according to a third-party data model) and zip the data, and augment the metadata using a third-party resource, Global Biodiversity Information Facility (GBIF)[4].

This project provides a model that other institutions can follow when dealing with homogenous, "time release" data from a campus partner. Specifically, it will describe the first two phases of the UMMZ CT Scanning Project partnership in detail and provide some discussion on the planned connections with the MorphoSource repository. Particular attention will be given to data generation, metadata requirements, workflows, code development, and lessons learned.

Project Background

Before approaching the Library, UMMZ had been storing raw data on external hard drives with no replication or backups in place. The IT department supporting the museum was not resourced to store or share publicly 30-40TB of data. In addition to these immediate storage and sharing needs, UMMZ was also looking for a long-term preservation and multifaceted discovery solution that would mint DOIs and share metadata with iDigBIO, GBIF and MorphoSource to complete the data management requirements of the NSF grant.

Implementing the oVert project required a significant deviation from the typical data deposits received by the DBRRDS team. Most deposits into Deep Blue Data are "one-shot" deposits where "return" depositors are infrequent and their deposits can vary greatly. DBD also has a mediated self-deposit model that allows researchers to fill in metadata and upload files smaller than 2GB on their own.

For the oVert project to succeed, DBRRDS had to develop a strong partnership with UMMZ as the data producers and the domain experts. These kinds of deep partnerships between libraries, researchers and other campus units such as IT are becoming increasingly important as data sharing and preservation requirements grow. Libraries are being asked not just to provide services, but to develop infrastructure to address researcher needs. For example, the New York University library and child learning and development community came together to develop Databrary, a disciplinary repository specifically designed for research videos. In addition to data description being informed by disciplinary knowledge, Gordon et al (2015) noted "… developing successful data repositories also requires new practices to manage workflows involving technology and metadata creation."

DBD is an institutional repository largely unknown to the 3D data or natural history museum community. To bolster discovery of the UMMZ datasets, links to the works will be added to the GBIF occurrence records via an aggregator attached to the Specify database. In addition, the third phase of this project entails sharing the metadata with MorphoSource to further extend discovery of the data sets. The diversity of storage and discovery, DBD's ability to handle large datasets, and providing full details about the images (how they were created, the specimen, terms of use, etc.) align with most best practices for 3-D morphological data (Davies et al., 2017)

[3] Specify: https://www.sustain.specifysoftware.org/
[4] GBIF: https://www.gbif.org/en/

Data Generation

Specimen scanning and preprocessing

Researchers at UMMZ used x-ray computed tomography (CT) scanning in the specimen digitization process to generate high-resolution anatomical data that will be represented as both 2D image stacks and 3D volumes and surfaces. The resulting data will provide unique "3-D visual replica[s] that can be virtually dissected, layer by layer, to expose cross-sections and internal structures" of specimens (Erickson, 2017). Figure 2 presents the scanning process in simplified form.

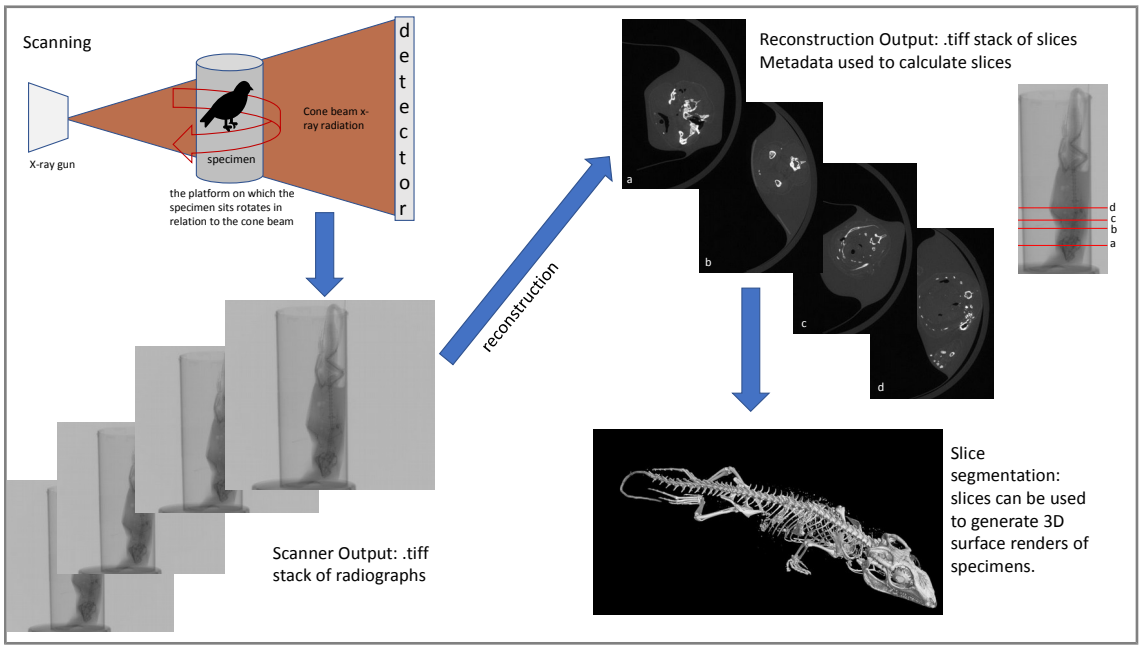


Figure 2. UMMZ simplified specimen scanning process. (Ramon Nagesan, personal communication).

The UMMZ scanning process produces raw data comprised of hundreds of radiograph images (x-rays similar to those from a radiologist) in stacks of TIFFs. Next, Nikon CT Pro 3D software is used to reconstruct the radiograph images into TIFF stacks of "slices" to build 3-D reconstructions.[5] These 3-D reconstructions are generated in data analysis software such as VGStudio Max[6], Dragonfly, Avizo, etc. The *.ply file type allows researchers to visualize the data and do analyses. Files for creating the 3-D reconstructions using VGStudio Max are included in the dataset. UMMZ makes the data available to DBRRDS in a "time released" fashion, releasing 30-50 data packages at a time. Each package totals 12–18 GB. For the project as a whole, these packages will potentially amount to several terabytes (TB).

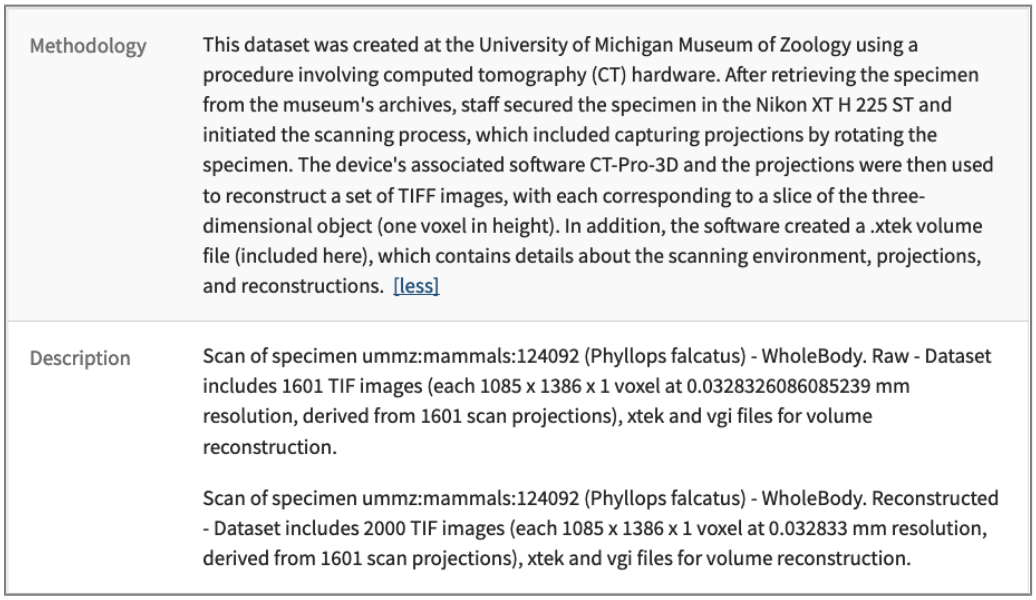
Metadata requirements

As with any data project data, description is a major component. Fortunately, this project has datasets with metadata that is consistent from dataset to dataset; only the number of TIFFs and

[5] In this article, "raw" data refer**s** to these x-ray image TIFFs and "reconstructed" data refer**s** to the "slice" TIFFs.
[6] VGStudio Max: https://www.volumegraphics.com/en/products/vgstudio-max.html

the specimen-specific values (ID, scientific name, etc.) vary. This consistency allows for automated metadata creation with a Python script that reads text files generated from scanning and uses the GBIF API (Chamberlain & Boettiger, 2017) to fill in the missing information about the specimen itself. As DBRRDS was not looking to automate metadata creation in general, the team developed a script unique to this project rather than using an existing application (Park & Brenza, 2015; Prabhune et al., 2015).

It has been noted that an image stack alone will not contain all the information necessary to make full use of the data (Davies, Rahman et al. 2017). Therefore, to expand the utility of the oVert data, part of the work level metadata includes information about the scanning process, the resolution (number of voxels—or 3-D pixels) and projections. Metadata values related to the scanning process were retrieved from the *.xtekct and *.xtekVolume text files (which are automatically generated as part of the scanning process): VoxelSizeX, VoxelsX, VoxelsY, VoxelsZ, and number of projections. Information about the scanning device and software used is displayed in the record for the dataset in DBD through the Methodology field. Specifics about the scan itself are displayed in the Description field as shown in Figure 3.



Methodology	This dataset was created at the University of Michigan Museum of Zoology using a procedure involving computed tomography (CT) hardware. After retrieving the specimen from the museum's archives, staff secured the specimen in the Nikon XT H 225 ST and initiated the scanning process, which included capturing projections by rotating the specimen. The device's associated software CT-Pro-3D and the projections were then used to reconstruct a set of TIFF images, with each corresponding to a slice of the three-dimensional object (one voxel in height). In addition, the software created a .xtek volume file (included here), which contains details about the scanning environment, projections, and reconstructions. [less]
Description	Scan of specimen ummz:mammals:124092 (Phyllops falcatus) - WholeBody. Raw - Dataset includes 1601 TIF images (each 1085 x 1386 x 1 voxel at 0.0328326086085239 mm resolution, derived from 1601 scan projections), xtek and vgi files for volume reconstruction. Scan of specimen ummz:mammals:124092 (Phyllops falcatus) - WholeBody. Reconstructed - Dataset includes 2000 TIF images (each 1085 x 1386 x 1 voxel at 0.032833 mm resolution, derived from 1601 scan projections), xtek and vgi files for volume reconstruction.

Figure 3. Detail from Deep Blue Data "Work Description" showing Methodology and Description.

Deep Blue Data uses Dublin Core[7] as the basis of its work description metadata while UMMZ uses Darwin Core[8] for its metadata standard. Rather than create new fields to accommodate this difference, values for these metadata fields were integrated throughout each deposit's "Work Description." The Darwin Core triple in the initial folder name provided by UMMZ ("ummz-mammals-124092_Phyllops-haitiensis [2018-09-17 10.55.29]") was used as the basis for the specimen metadata which is acquired using the GBIF API. As shown in Figure 4, specimen metadata is displayed across several fields: Title and Description (as the scientific name), Keyword (as kingdom, phylum, class, order, family, scientific name, and GBIF occurrence ID), and in Citations to Related Material (a link to the GBIF record).

[7] Dublin Core: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[8] Darwin Core: https://dwc.tdwg.org/terms/

Keyword	Animalia Chordata Mammalia Chiroptera Phyllostomidae Phyllops falcatus 1987339099 computed tomography X-ray 3D
Date coverage	2018-09-17
Citations to related material	For more information on the original UMMZ specimen, see: https://www.gbif.org/occurrence/1987339099

Figure 4. Detail from Deep Blue Data "Work Description" showing Keyword and Citations to related material.

Workflows

Getting batches of datasets, 10 – 50 at a time, to be ingested at regular intervals was a new experience for DBRRDS. New technical and human processes to handle UMMZ's data needed to be in place for the ingest to happen successfully. UMMZ had a defined workflow for scanning and preprocessing the specimens and DBRRDS had a defined process for depositing individual datasets. These two workflows had to be combined and additional subprocesses had to be developed for the overall process to work in harmony. Meetings with active whiteboard diagramming and regular updates between the teams led to the combined scanning-data deposit process. Figure 4 provides an overview of the resulting process.

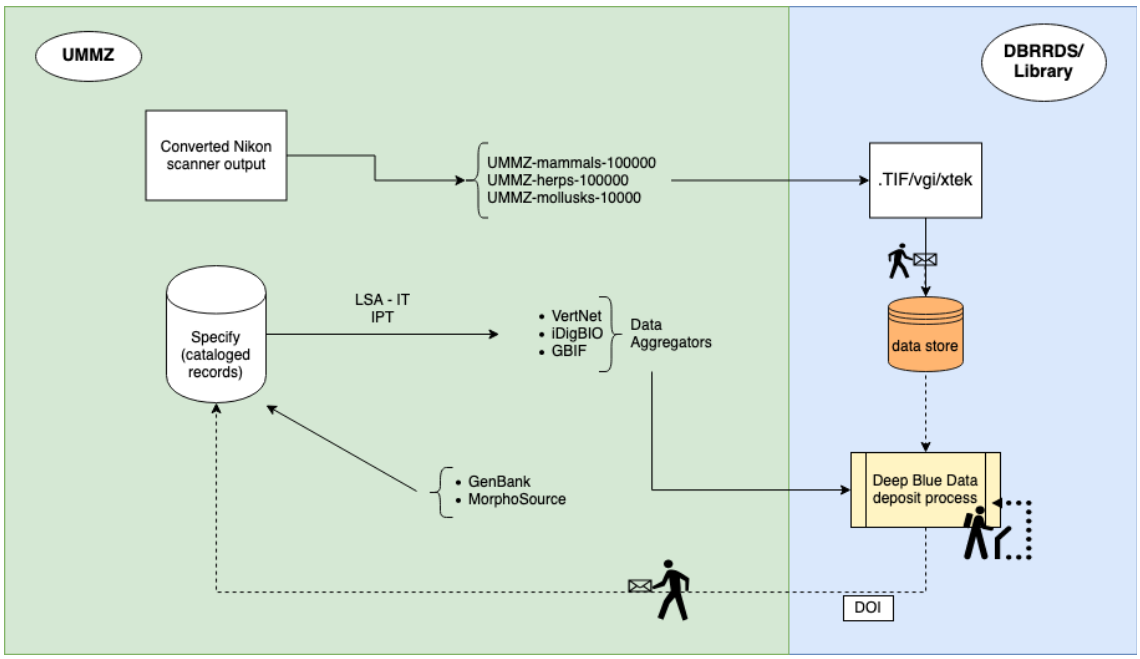


Figure 1. Overview of UMMZ to DBRRDS workflow, steps 1-3 and 9 (below).

UMMZ scanning and DBD workflow

1. UMMZ scans the specimen

2. Output from a proprietary Nikon format is converted to TIFF
3. UMMZ puts data in lab server transfer location
4. DBRRDS copies data from transfer location to library IT server (Figure 5 illustrates steps 4-8)
5. DBRRDS runs reorganization Python script
6. DBRRDS runs Python script to gather metadata, zip folders and create *.yml files
7. DBD system administrator runs Ruby rake task to create new deposits in DBD per *.yml files and upload data files.
8. DBD system administrator runs report to capture DOIs of new deposits and sends them to UMMZ.
9. UMMZ updates Specify which then updates iDigBIO, VertNet and GBIF with DOIs pointing back to specimen data in Deep Blue Data.

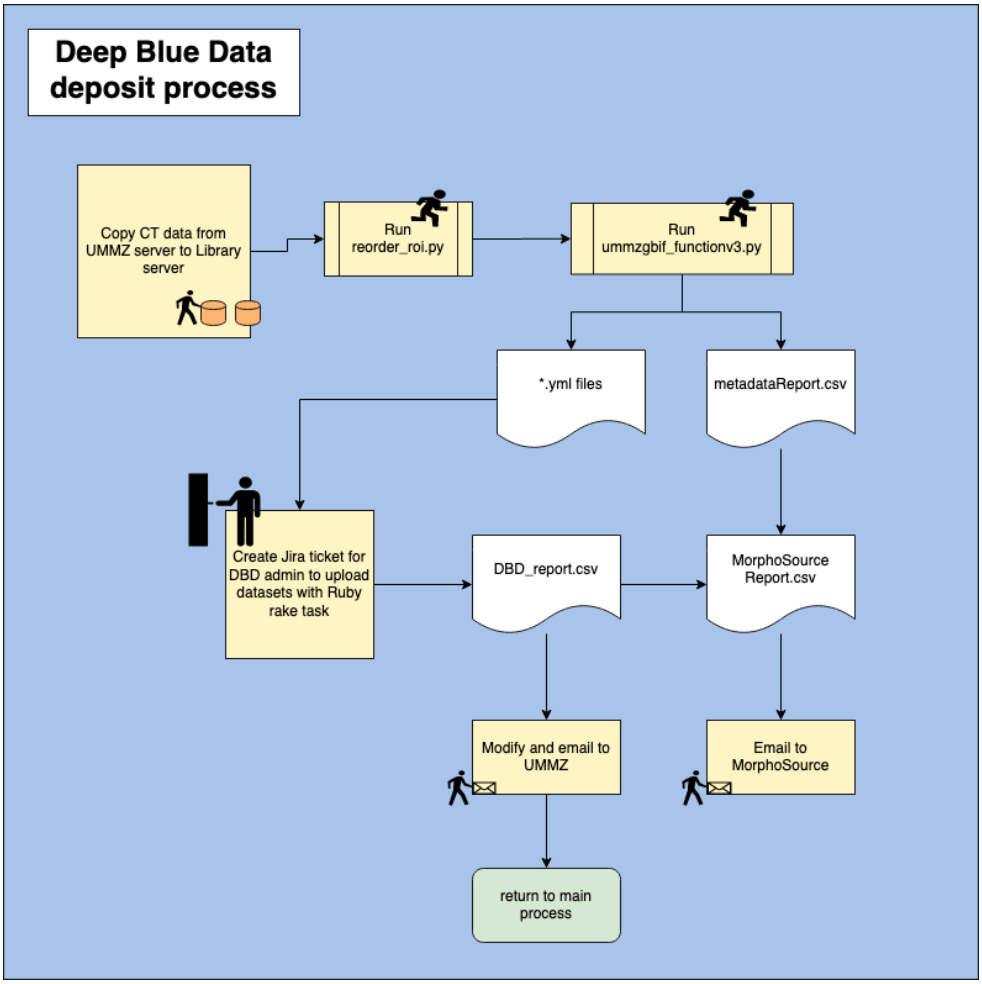


Figure 5. Detail of DBRRDS specific workflow, steps 4-8.

DBRRDS staff had to be granted access to the UMMZ lab server to copy the batch data to the library server for pre-ingest processing. To provide a data safety net for problems during

processing or ingest, the project team also had to reach an agreement on a retention schedule for data on the UMMZ server after retrieval by DBRRDS.

Code Development

The author created Python scripts to process oVert data further and prepare it for ingest to Deep Blue Data. This processing included reorganizing the files in each data package and augmenting the metadata primarily with specimen information from GBIF. A metadata record (Deep Blue Data work metadata), *.yml file, had to be created for each data package. The code for these processes is comprised of two scripts both of which are available on GitHub[9]. This code processes the data in batches, folders of data, as supplied by UMMZ.

"reorder_roi.py[10]" script

This script reorganizes the dataset such that within each top-level folder the raw CT scan TIFF stack and related files and the reconstructed CT scan TIFF stack and related files are placed into folders. The folders also need to be renamed to indicate the "region of interest" (ROI) that was scanned. For example, in Figure 6 the top-level folder "ummz-mammals-124092_Phyllops-haitiensis [2018-09-17 10.55.29] does not have "skull" or another ROI in the string. The ROI is therefore assumed to be "whole body" as shown in Figure 6. Reptiles and amphibians are frequently scanned by ROI such as Skull and WholeBody.

Everything in the "ummz-mammals-124092_Phyllops-haitiensis_01" is "Reconstructed" data because the folder, "[vg-project] ummz-mammals-124092_Phyllops-haitiensis," contains all the information VGStudio Max needs to create a 3-D reconstruction (see Figure 7) and the TIFF are the "slice" data used in the 3-D reconstruction process. As such, the folder ending in "_01" was renamed with "Recon" and everything then else had to be moved into a new folder for "Raw" data as seen in Figure 7.

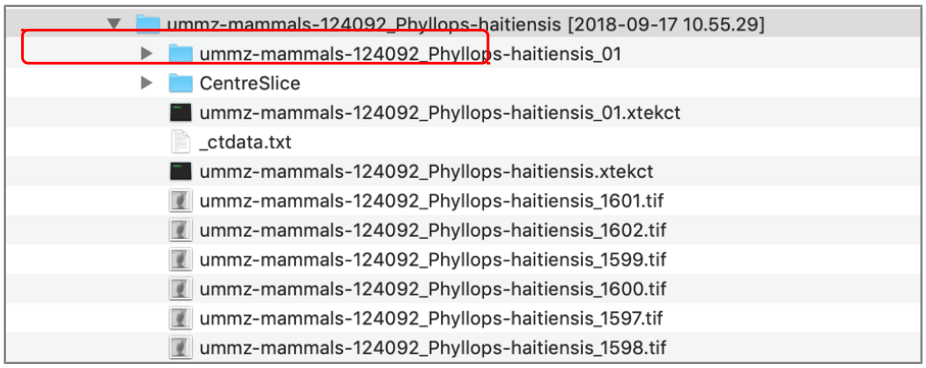


Figure 6. Initial dataset organization from UMMZ.

[9] GitHub: https://github.com/mutanthumb/ummzgbif
[10] Reorder_roi.py: https://github.com/mutanthumb/ummzgbif/blob/master/reorder_roi.py

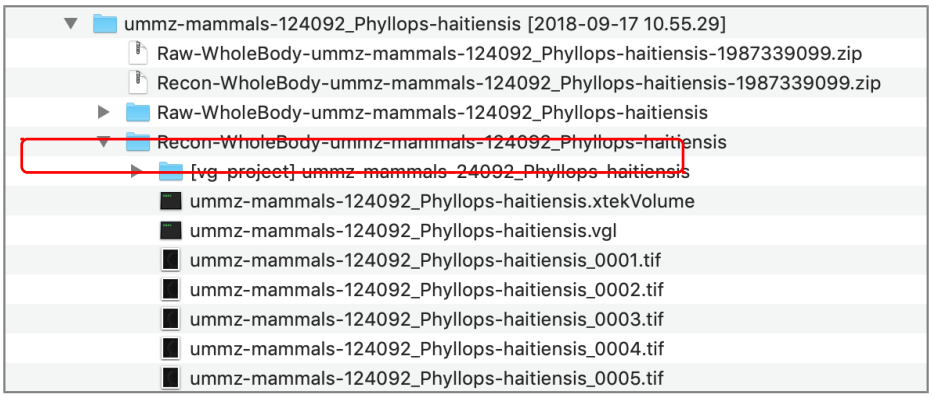


Figure 7. Dataset after folder reorganization.

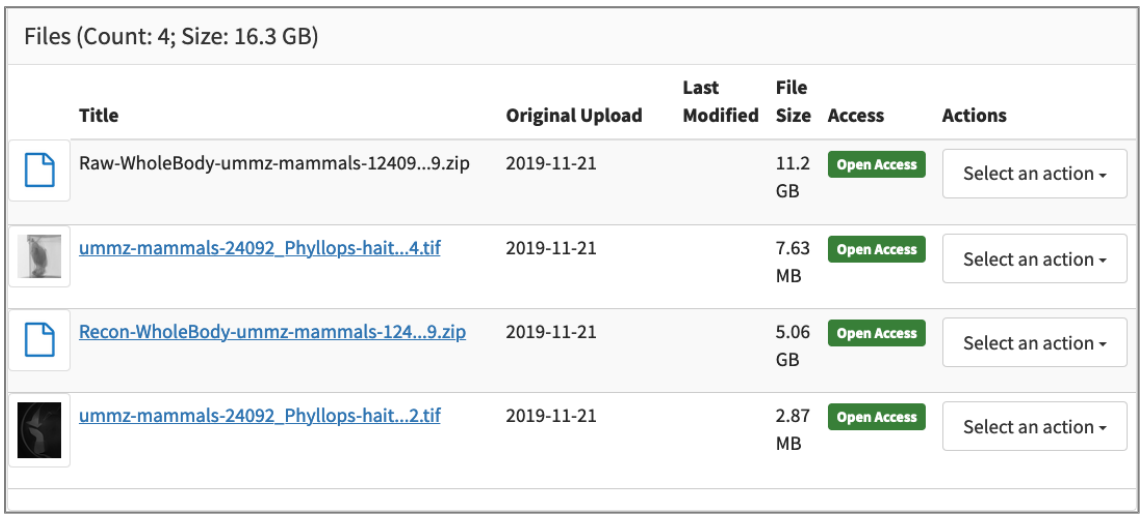


Figure 8. Final file organization in Deep Blue Data. https://deepblue.lib.umich.edu/data/concern/data_sets/3b5918681?locale=en.

"ummzgbif_functionv3.py[11]" script

This script begins with parsing the folder name provided by UMMZ for example, ummz-mammals-24092_Phyllops-haitiensis [2018-09-17 10.55.29]. The script then sends those parsed parts "institution code" (UMMZ), "collection code" (mammals), and the "catalog number" (124092) to GBIF's API URL and gets the JSON back.

```
https://api.gbif.org/v1/occurrence/search?
catalog_number=124092&collection_code=mammals&institution_code=ummz
```

Error handling is included in this section to catch issues with records not yet in existence or other problems. For example, after parsing the folder name and running it through the script, it returned an error. A typo in the catalog number "24092" produced the error; the number should have been "124092". The initial catalog number can still be seen in the *.xtekct file in Figure 8. Next, the script parses the JSON results to get the "key" which is the GBIF occurrence ID (highlighted in yellow).

```
{"offset":0,"limit":20,"endOfRecords":true,"count":1,"results":
[{"key":1987339099,"datasetKey":"6d2cfc0a-9903-40b8-802b-403398218e4a
```

[11] Ummzgbif_functionv3.py: https://github.com/mutanthumb/ummzgbif/blob/master/reorder_roi.py

This key is used with the "/fragment" URL:

https://api.gbif.org/v1/occurrence/1987339099/fragment

The "fragment" URL returns JSON results that need to be parsed for Darwin Core values (highlighted in yellow):

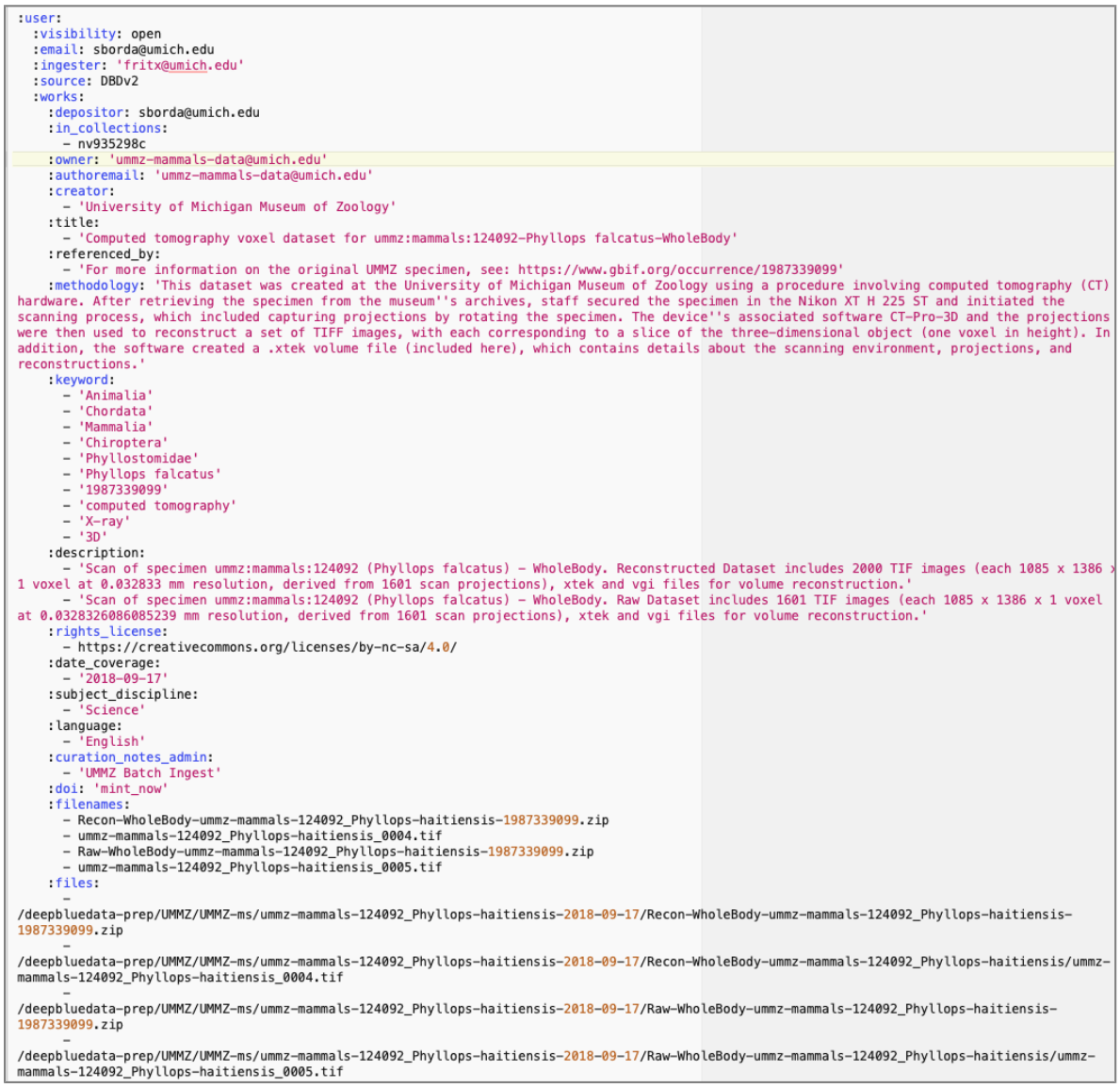
{"basisOfRecord":"PreservedSpecimen","catalogNumber":"124092","class":"Mammalia","collectionCode":"mammals","continent":"NORTH AMERICA","coordinateUncertaintyInMeters":"1.0000000000","country":"HAITI","county":null,"day":"07","decimalLatitude":"18.3000000000","decimalLongitude":"-73.9333300000","establishmentMeans":"Native","eventDate":"01-07-1975","extensions":{"dwc:ResourceRelationship":[{"relatedResourceID":"urn:catalog:UMMZ:Mammals:124092","relationshipAccordingTo":"VertNet","relationshipEstablishedDate":"2019-01-14 09:47:14.0","relationshipOfResource":"sameAs","resourceID":"d828a605-c1b1-42ac-8ef2-ac2a601e3c07"}]},"family":"Phyllostomidae","fieldNumber":null,"genus":"Phyllops","geodeticDatum":"WGS84","georeferenceProtocol":"MaNIS georeferencing guidelines","georeferenceRemarks":null,"georeferenceSources":"Alexandria Digital Library Gazetteer","georeferenceVerificationStatus":"unverified","georeferencedBy":"Lucy Tran","georeferencedDate":null,"higherClassification":"Chiroptera Phyllostomidae Phyllops","higherGeography":"HAITI, SUD","id":"d828a605-c1b1-42ac-8ef2-ac2a601e3c07","infraspecificEpithet":null,"institutionCode":"UMMZ","kingdom":"Animalia","lifeStage":null,"locality":"PAILLANT,6KM SW MIRAGUANE","modified":"2017-4-04","month":"01","nomenclaturalCode":"ICZN","occurrenceID":"d828a605-c1b1-42ac-8ef2-ac2a601e3c07","occurrenceRemarks":null,"occurrenceStatus":"Present","order":"Chiroptera","otherCatalogNumbers":null,"phylum":"Chordata","preparations":"FLUID-ALCOHOL - 1","recordNumber":"1353","recordedBy":"Klingener, D.","reproductiveCondition":null,"scientificName":"Phyllops falcatus","sex":"FEMALE","specificEpithet":"falcatus","stateProvince":"SUD","taxonRank":"species","taxonomicStatus":"Current","typeStatus":null,"verbatimCoordinateSystem":"decimal degrees","verbatimElevation":null,"verbatimEventDate":"7 JANUARY 1975","verbatimLatitude":"18.3","verbatimLocality":"HAITI: SUD: CO.; PAILLANT,6KM SW MIRAGUANE; 18.3, -73.93333;","verbatimLongitude":"-73.933329999999998","year":"1975"}

The "raw" and "recon" folders are zipped for easier ingest into Deep Blue Data. The *.xtekct or *.xtekVolume files are opened and read for scanning metadata, such as voxel sizes (Figure 9) and number of projections.

```
1   [XTekCT]
2   Name=ummz-mammals-24092_Phyllops-haitiensis
3   InputSeparator=_
4   OutputSeparator=_
5   InputFolderName=
6   OutputFolderName=ummz-mammals-24092_Phyllops-haitiensis_01
7   VoxelsX=1085
8   VoxelsY=1386
9   VoxelsZ=2000
10  VoxelSizeX=0.0328326086085239
11  VoxelSizeY=0.0328326086085239
12  VoxelSizeZ=0.0328326086085239
13  OffsetX=-0.0463221515217214
```

Figure 9. Detail of *.xtekct file showing Voxel sizes.

Finally, the metadata is combined with standard information as well as file names and locations to create a *.yml for the deposit description as shown in Figure 10.



```
:user:
  :visibility: open
  :email: sborda@umich.edu
  :ingester: 'fritx@umich.edu'
  :source: DBDv2
  :works:
    :depositor: sborda@umich.edu
    :in_collections:
      - nv935298c
    :owner: 'ummz-mammals-data@umich.edu'
    :authoremail: 'ummz-mammals-data@umich.edu'
    :creator:
      - 'University of Michigan Museum of Zoology'
    :title:
      - 'Computed tomography voxel dataset for ummz:mammals:124092-Phyllops falcatus-WholeBody'
    :referenced_by:
      - 'For more information on the original UMMZ specimen, see: https://www.gbif.org/occurrence/1987339099'
    :methodology: 'This dataset was created at the University of Michigan Museum of Zoology using a procedure involving computed tomography (CT)
hardware. After retrieving the specimen from the museum''s archives, staff secured the specimen in the Nikon XT H 225 ST and initiated the
scanning process, which included capturing projections by rotating the specimen. The device''s associated software CT-Pro-3D and the projections
were then used to reconstruct a set of TIFF images, with each corresponding to a slice of the three-dimensional object (one voxel in height). In
addition, the software created a .xtek volume file (included here), which contains details about the scanning environment, projections, and
reconstructions.'
    :keyword:
      - 'Animalia'
      - 'Chordata'
      - 'Mammalia'
      - 'Chiroptera'
      - 'Phyllostomidae'
      - 'Phyllops falcatus'
      - '1987339099'
      - 'computed tomography'
      - 'X-ray'
      - '3D'
    :description:
      - 'Scan of specimen ummz:mammals:124092 (Phyllops falcatus) - WholeBody. Reconstructed Dataset includes 2000 TIF images (each 1085 x 1386 x
1 voxel at 0.032833 mm resolution, derived from 1601 scan projections), xtek and vgi files for volume reconstruction.'
      - 'Scan of specimen ummz:mammals:124092 (Phyllops falcatus) - WholeBody. Raw Dataset includes 1601 TIF images (each 1085 x 1386 x 1 voxel
at 0.0328326086085239 mm resolution, derived from 1601 scan projections), xtek and vgi files for volume reconstruction.'
    :rights_license:
      - https://creativecommons.org/licenses/by-nc-sa/4.0/
    :date_coverage:
      - '2018-09-17'
    :subject_discipline:
      - 'Science'
    :language:
      - 'English'
    :curation_notes_admin:
      - 'UMMZ Batch Ingest'
    :doi: 'mint_now'
    :filenames:
      - Recon-WholeBody-ummz-mammals-124092_Phyllops-haitiensis-1987339099.zip
      - ummz-mammals-124092_Phyllops-haitiensis_0004.tif
      - Raw-WholeBody-ummz-mammals-124092_Phyllops-haitiensis-1987339099.zip
      - ummz-mammals-124092_Phyllops-haitiensis_0005.tif
    :files:
      -
/deepbluedata-prep/UMMZ/UMMZ-ms/ummz-mammals-124092_Phyllops-haitiensis-2018-09-17/Recon-WholeBody-ummz-mammals-124092_Phyllops-haitiensis-
1987339099.zip
      -
/deepbluedata-prep/UMMZ/UMMZ-ms/ummz-mammals-124092_Phyllops-haitiensis-2018-09-17/Recon-WholeBody-ummz-mammals-124092_Phyllops-haitiensis/ummz-
mammals-124092_Phyllops-haitiensis_0004.tif
      -
/deepbluedata-prep/UMMZ/UMMZ-ms/ummz-mammals-124092_Phyllops-haitiensis-2018-09-17/Raw-WholeBody-ummz-mammals-124092_Phyllops-haitiensis-
1987339099.zip
      -
/deepbluedata-prep/UMMZ/UMMZ-ms/ummz-mammals-124092_Phyllops-haitiensis-2018-09-17/Raw-WholeBody-ummz-mammals-124092_Phyllops-haitiensis/ummz-
mammals-124092_Phyllops-haitiensis_0005.tif
```

Figure 10. Sample YML file created by ummzgbif_functionv3.py

Currently in DBD, DOIs are not automatically assigned to new works as they are created. In order to trigger the creation of a DOI at the time of ingest, "mint_now" was used as the value for the ":doi:" field in the *.yml file. The collection ID is also included in the *.yml file to have the works automatically assigned the appropriate collection upon creation.

Lessons Learned

In the process of developing and implementing the oVert project, team members made the following important discoveries:

- Determine the sustainability of the source for specimen metadata, to use iDigBIO, or GBIF. The project team decided to switch from iDigBIO to GBIF mid-project because of stability concerns, which meant adjustments to the script that captured the metadata.

- Some issues are found only when code is applied. For example, noticing that GBIF records were not showing the second epithet of the scientific name → had to change the GBIF API end-point to /Fragment.

```
record_results = requests.get(gbif_baseurl + 'occurrence/' +
str(item['key']) + '/fragment')
```

- Thoroughly review test uploads into DBD for completeness. For example, the "Citations to related materials" field from the *.yml file was not being uploaded to the work.

- Code can always benefit from additional error handling.

- Confirm the data model at the very beginning. The project team spent a lot of time working through various data models.

Next Steps

The oVert initiative at the University of Michigan is an ongoing project for which the pilot phase is ending. The following changes will be made to improve the implementation of future phases:

- The number of manual activities will be reduced to resolve issues of scale. Scheduled cron jobs will be created to copy the data and delete it from the UMMZ server. A batch script will be written to run the reorganization and metadata scripts serially. In addition, the version of Python on the server will be upgraded from 2.7 to 3.7 so the scripts can be run from the server instead of a PC in the DBRRDS office.

- Plale et al., (2013) note that "Even if institutional repositories removed major obstacles to data submission and researchers began to submit their data, the view of data would be a fragmented one; a researcher would have to search repositories one by one to find relevant data." To address this problem, the oVert team will use a "discovery layer." As the TIFF stack data gets added to Deep Blue Data, the work and scanning metadata will be shared with MorphoSource — a 3D data aggregator and disciplinary repository.

Conclusion

The oVert initiative is already expanding the reach and utility of the library's data management services. It can now demonstrate a successful template for engagement and partnership with the museums and other groups at the University of Michigan with similar data issues: homogenous data, storage constraints and sharing needs.

Through the oVert partnership, the library is gaining ready access to domain knowledge for checking metadata that is frequently lacking in single data deposits. It is also gaining a foothold on the larger museum landscape at the University of Michigan. For example, members of the UM Museum of Natural History paleontology team have been involved in some of library's project meetings to see how their data might fit a similar workflow. In addition, the partnership with MorphoSource allows DBRRDS to expand the discovery capability of Deep Blue Data to a specific disciplinary audience. This case study is a model that other institutions can follow when dealing with homogenous, "time release" data from a campus partner.

Acknowledgments

This would not be a project without the time and effort put in by Ramon Nagesan and Dr. Cody Thompson from UMMZ as well as the guidance of Jake Carlson from DBRRDS.
References