MorphoSource's approach to describing, preserving and managing complex representational data

Julie M. Winchester / Product Manager & Lead Dev.
Doug M. Boyer / MorphoSource Director
Tim Ryan / Professor PSU Anthropology
Tim McGeary / Duke Associate Librarian of IT

Part II: Integrating MorphoSource with other resources
• 3D data can be viewed as a microcosm for representational data more broadly
  • Direct representation of object, but...
  • High metadata load
  • Lack of easy software support
  • No best practices for linkages
  • Data produced by many people in many ways

• Solutions for 3D data can possibly be instructive for representational data more broadly
• Similarly, MorphoSource reflects challenges of representational data repositories
  • Supporting challenging type of data (3D)
  • Providing platform for individuals and groups
  • Archiving and serving access to data
  • Demonstrating value of data represented in archive

• Describing MorphoSource’s journey and design to both seek feedback for self-improvement and hopefully say something about broader issues at play
• MorphoSource 1 (MS1) Development Timeline
  • How was the site built?
  • What stressors were experienced?

• MorphoSource 2 (MS2) State of the Refactor
  • Development
  • Tour of features

• Integrating MorphoSource with Other Resources
Getting Started

Find & Download Datasets

Useful Info
- About MorphoSource
- Information for Users
- Information for Contributors
- Staff
- User Guide

Recently Published
A new genus and species of clingfish from the Rangitahoe Kermadec Islands of New Zealand

See all project specimens
Read the published article

Welcome

MorphoSource is a project-based data archive that allows researchers to store and organize, share, and distribute their own 3D data. Furthermore, any registered user can immediately search for and download 3D morphological data sets that have been made accessible through the consent of data authors.

The goal of MorphoSource is to provide rapid access to as many researchers as possible, large numbers of raw microCT data and surface meshes representing vouchered specimens.

File formats include tiff, dicom, standford ply, and stl. The website is designed to be self-explanatory and to assist you through the process of uploading media and associating it with meta data. If you are interested in using the site for your own data but have questions about security or anything else contact the site administrator. Otherwise please download whatever data you need and check back frequently to see what's new.
MorphoSource 1 Development Timeline

Time

Number of code contributions

Initial Development

Maintenance, Scaling

Shift to Internal Dev, Triage
• February 13, 2013: First Code Contribution

• MS1 built on Collective Access
  • Open-source software for managing and publishing museum and archival collections
  • jQuery/PHP
  • Significantly customized to support 3D and media data model
  • Proof of concept
• New Features
  • Media Cart
  • DOIs

• Scaling and Stressors
  • “Publication Packet” model
  • Media ownership
  • More institutional users and use cases
  • Aging UI
Aging technical stack (and UI) makes improvements more difficult

- Tech debt: Code accretion, lack of modularity, loose separation of powers

- Unable to update core framework

- jQuery/PHP: modern “web application” solutions difficult to impossible

- Server-side Meshlab a nightmare
2017-Now: Internal Dev & Triage

- Shifting to internal development (funding!)
  - More code contributions, more regular small-scale feature updates
  - New features: project dashboard UI revamp, file upload interruption protection, automated data reporting to institutions...

- Acceleration of scaling
  - oVert

- 2018: Beginning development of MS2
• MS2 is a Hyrax application
  • Open-source Samvera-powered repository solution
  • Ruby on Rails
  • Large, active community ecosystem
  • Relatively fresh tech stack
• Why Hyrax?
  • Standardize with existing Library repositories
  • Background job queueing
  • Robust multiple file upload, cloud plug-ins
  • Built-in derivative generation
  • User and group support
  • Storing records (media) in multiple collections (projects)
  • Modular design with solid design principles
  • Universal Viewer (+IIIF) for record preview by default
  • Google analytics
  • More suitable for automated deployment
MorphoSource 2 Development

• Starting with...
  • MS1
  • Appreciation for future needs
  • Vanilla Hyrax (significant customization)
    • Minimal working back-end AND front-end

• Rough development procedure (cyclical!)
  • Identify user needs via user stories
  • Implement back-end features
  • Make features accessible via front-end

• Logistics
  • Agile Sprints
  • Jira Issue Tracker
MS2 Data Model

Specimen
Occurrence ID: a2770...

Media
MS ID: M1770

Media
MS ID: M1771

Entity

Event

PREMIS

PRESERVATION METADATA
MAINTENANCE ACTIVITY
Specimens and media organized by “project” containers, by institution or collection, by tagging.
Significant processing without intermediate media

1. Smoothing
2. Simplification
3. Hole Filling

Significant processing with intermediate media

1. Smoothing
2. Simplification
3. Hole Filling
We need to gather some initial information about your media. Please select all that apply concerning raw or derived status. Please note for CT or MRI media, we consider reconstructions to be derived from parent projection media, even though in most cases we suggest users do not upload projections.

- My media is not derived from any parent source media.
- My media is derived from some parent media. Select this even if you do not plan to upload the parent media.
- There is media already on MorphoSource that is derived from my media.

**THIS ACCURATELY DESCRIBES MY MEDIA**
We assume your media file(s) represents a physical object. Is this physical object better described as a biological specimen or a cultural heritage object?

- BIOLOGICAL SPECIMEN OBJECT
- CULTURAL HERITAGE OBJECT
Search for a biological specimen already in MorphoSource or iDigBio, or select Create New Biological Specimen.

- **Occurrence ID**: Lorem ipsum
- **Institution Code**: Lorem ipsum
- **Collection Code**: Lorem ipsum
- **Catalog Number**: Lorem ipsum
- **Genus**: Lorem ipsum
- **Species**: Lorem ipsum

[SEARCH] [CREATE NEW BIOLOGICAL SPECIMEN]
If this object is managed by an organization, either select it from the list or create a new organization record if it is not currently present in the list. If the object is not managed by an organization, select No Institution.
Select a taxonomy to associate with this biological specimen by searching below. Or select Create New Taxonomy.

Enter taxonomy keywords here

[SELECT TAXONOMY] [CREATE NEW TAXONOMY]
Finish creating a new biological specimen by filling in the following fields and selecting Create Biological Specimen.

**SPECIMEN IDENTIFIERS AND EXTERNAL LINKS**

- **Voucher:**
  - **Id:** Select
  - **Reg:**

- **Institution Code:**
  - **Id:** Select
  - **Reg:**

- **Collection Code:**
  - **Recommended:**
  - **Collection Code:**

- **Catalog Number:**
  - **Recommended:**
  - **Catalog Number:**

- **Occurrence ID:**
  - **Recommended:**
  - **Occurrence ID:**

- **Other Identifier(s):**
  - **Other identifier:**

- **External Object URL:**
  - **External Object URL:**

**GENERAL DETAILS**

- **Description:**

- **Sex:**
  - **Select:**

- **Is Type Specimen:**
  - **Select:**

- **Source Of Record:**
  - **IdDigBio:**

- **Data Created:**
  - **Date Created:**

- **Date Uploaded:**
  - **February 19, 2019**

**COLLECTIONS PHYSICAL OBJECT BELONGS TO**

- **Search Collections**
  - **American Museum of Natural History**
  - **Mammalogy Collections**
  - **Cercopliocerids of West Africa**
  - **Winchester 2016: Dental topography of extant cercopliocerids**

**PHYSICAL OBJECT TAGS**

- **Enter Tag**

- **Lorem ipsum**
  - **Dolor**
  - **Sit amet**
  - **Sapiente**

- **Nunc amet dolor ac odio**

**BIBLIOGRAPHIC CITATIONS FOR THIS PHYSICAL OBJECT**

- **Add DOIs to associate citations to physical object**

- **[Add DOIs]**

- **[Add DOIs]**

- **[Add DOIs]**
MS2 Media Submission

Finish creating new media by filling in the following fields and selecting Create Media.

GENERAL DETAILS
- Media Modality
- Media Type
- Element Or Part Of Object
- Side
- Orientation
- Short Description
- Full Description

COLLECTIONS MEDIA BELONGS TO
- Search Collections
- American Museum of Natural History Mammalogy Collections
- Cercopithecoids of West Africa
- Winchester 2016: Dental topography of extant cercopithecoids

OWNERSHIP, PERMISSIONS, AND ACKNOWLEDGEMENTS
- Data Managed By
- Download Permission
- Usage License
- Copyright Statement
- Custom Terms Of Use
- Additional Usage Agreement URL
- Commercial Use Permitted

MEDIA TAGS
- Enter Tag
- Lorem ipsum
- Dolor
- Sit Amet
- Sapient etiam
- Nunc amet dolor ac odio

FILE OBJECT DETAILS
Please Note: If uploading a DICOM image stack, fields with the icon will be parsed automatically. Otherwise, enter this information.
Live Demo
The Division of Fossil Primates (DFP) is a research collection used to understand our evolutionary journey as primates.

**Organizations:**
- Title: Duke Lemur Center Division of Fossil Primates
- Institution Code: DPC
- Description: The Division of Fossil Primates (DFP) is a research collection used to understand our evolutionary journey as primates.
- Address: 1513 Broad Street
- City: Durham
- State/Province: NC
- Country: United States
- Institution Name: Duke University
- Collection Code:

**Team:**
- Title: Duke Lemur Center Division of Fossil Primates
- Institution Code: DPC
- Description: The Division of Fossil Primates (DFP) is a research collection used to understand our evolutionary journey as primates.
- Address: 1513 Broad Street
- City: Durham
- State/Province: NC
- Country: United States
- Institution Name: Duke University
- Collection Code:

**Collection Details:**
- Total Items: 4
- Size: 117 KB
- Location: Durham, NC
- Related URL: https://morpho.duke.edu/discover/division-of-fossil-primates/

**Works:**
- Title: Duke Lemur Center Division of Fossil Primates
- Date Added: 2020-01-21
- Visibility: Public
Integrating MorphoSource with other resources

Julie M. Winchester / Product Manager & Lead Dev.
Doug M. Boyer / MorphoSource Director
Tim Ryan / Professor PSU Anthropology
Tim McGeary / Duke Associate Librarian of IT
Linked Data (Berners-Lee, 2006)

- Use URIs (**uniform resource identifier**) as names for things
- Use HTTP URIs (AKA URLs) so people can look things up
- HTTP URIs should provide useful information when looked up
  - Machine readable
  - Standardized (RDF)
- Each URI resource should have links to other URI resources
- If resource is freely available, Linked Open Data

Tim Berners-Lee,
Inventor of the World Wide Web
Linked Open Data Cloud (www.lod-cloud.net)
• Does MorphoSource meet those goals?
  • Absolutely not!

• Part of the reason for this workshop, to move forward together

• Simple goals belie complexity and nuance of implementation
Integration Spectrum

Federation → Confederation
Integration Spectrum

Federation Characteristics
- Restricted data and use case domain
- Centralized authority, fewer groups for buy-in
- High ability to dictate uniformity
- Deep but narrow

Confederation Characteristics
- Broad and varied domain
- Decentralized authority, buy-in from many groups
- Community-wide solutions
- Need for translation/normalization
- Wide but shallow
Integration of 3D data resources
- Connecting existing data to/from MorphoSource
- Making MorphoSource 3D archival software more available

Integration with wider data world
- Connecting MorphoSource 3D data to...
  - Specimen repositories
  - Taxonomy resources
  - Other data archives
Integration of 3D data resources

• Metadata-only MorphoSource records
  • Data linkages

• MorphoSource-compatible 3D data archive instances
  • Automated deployment strategy
  • Turnkey solutions
  • Design complexity
Data Linkages

• Use case
  • MorphoSource user has data in web-accessible storage
  • Wants to make these data available in MorphoSource...
  • ... but for many possible reasons, does not want to re-store that data in MorphoSource

• Solution
  • Metadata-only MorphoSource records that point to remote resource
Requirements for basic version

• HTTP URL for individual media (mesh, CT stack, etc.)
• File binary accessible at HTTP URL
  • Ideally for routing users to download
  • At minimum, for derivative generation and fixity checks
• Machine readable metadata
• All of the above persistent and reliably accessible
• Ingest
  1. URL is specified for ingest
  2. MS GETs remote file and metadata
  3. Automated file characterization and derivative generation

• Fixity Checking
  4. On a regular basis, metadata and file integrity are verified
  5. Automated process for error reporting

• Access
  6. Users informed of remote nature
  7. Routed to remote file as appropriate
• The Future
  • Proof of concept for a standardized biodiversity repository media linkage system?

• Pros and cons
  • Pro: sustainability
  • Pro: community investment
  • Con: fragile
  • Con: bespoke, at least to start with
Use cases

Infrastructural resources > Funding resources

OR

Want to use MS data model and share data but store own data

OR

Want private/dark 3D data repository
MorphoSource Instances

- Advantages of this work outside the MorphoSource domain

- Automated deployment and/or containerization could be used to make biodiversity repository technology more portable and easy to use

- Enables use across a diversity of terrains, from cloud to local hardware
Containerization: Encapsulating a software application (or software application components) into its own environment, so it can run uniformly and consistently on any infrastructure.

Procedure

• Start with component application modules

• Wrap each module in a single container
  • Docker

• Connect containers
  • Docker compose

• Orchestrate containers in an automated way
  • Kubernetes
• XNHM has own MorphoSource instance

• Indexes instance-local media in MorphoSource.org

• Should XNHM be able to edit that media from MorphoSource.org?
  • Balkanization versus increasingly complex dynamics

• How about for researcher-scanned media of XNHM?
Examples of work previously completed

- Automated data reporting
- iDigBio specimen integration
- API
• Possible Future Directions
  • Finish MS2 (so we can move on to new things!)
  • Not just APIs, but standardized machine-readable metadata at URLs
  • Connecting media to specimen repositories
  • Better importation from specimen and taxonomy resources

• Priorities largely determined by feedback from this community!
MorphoSource

Acknowledgements


Internal Test Users: Mackenzie Nieto-Aguilar, Arianna Harrington, Ethan Fulwood, Bernadette Perchalski, Paul Morse, Gabe Yapuncich, Boyer Lab Group

MorphoSource/Duke Team

Primary Developers: Jocelyn Triplett, Simon Choy

Additional Development: Ryan Baumann, Jim Coble

DevOps: David Chandek-Stark, Jack Hill

Project Support: Will Sexton, Ed Gomes

Additional Duke Support: Charley Kneifel, Sophia Lafferty-Hess, Tracy Futhey, Mark DeLong

All MorphoSource contributors and users!