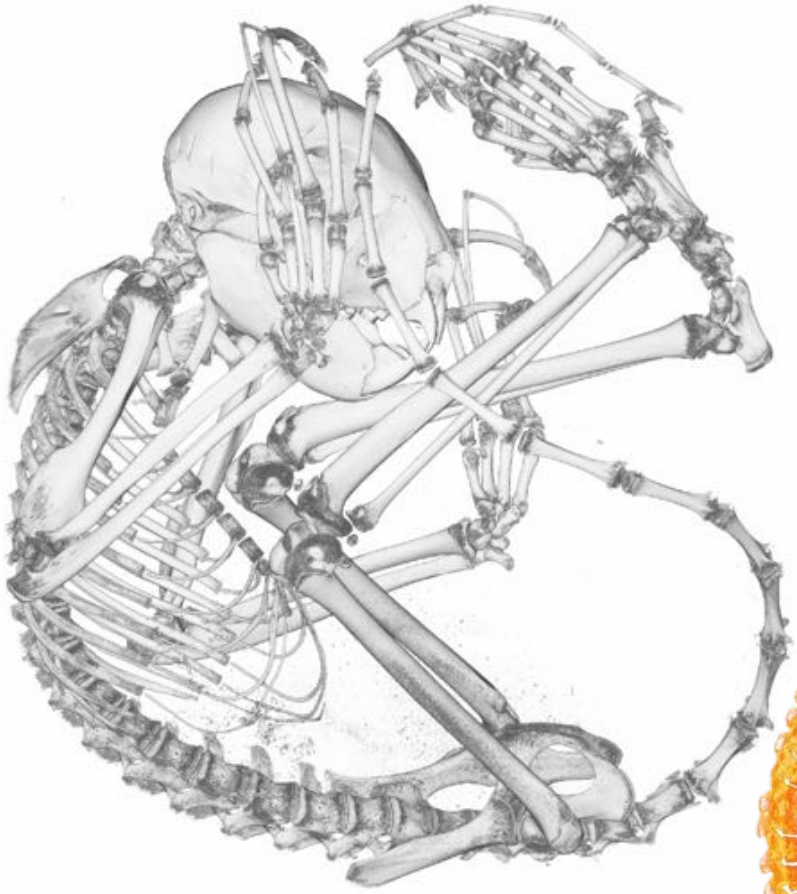
Digital Data in Paleontology. March 28, 2016

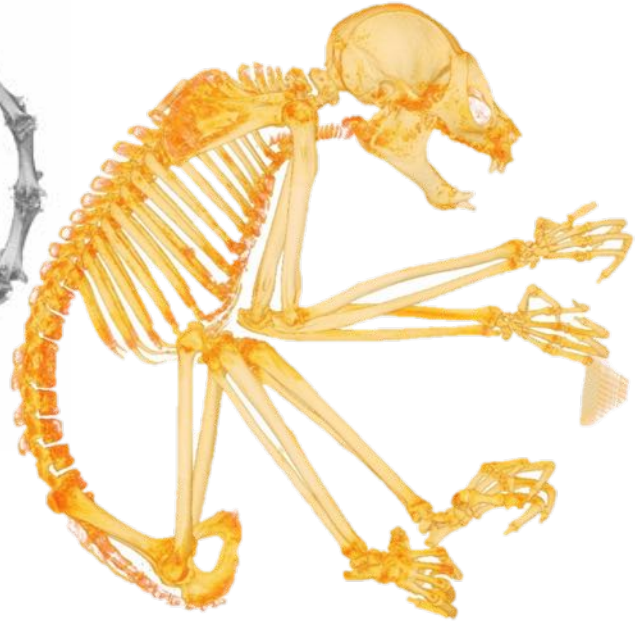
Managing Large Datasets
Archiving and Sharing 3D digital specimen data

Doug M. Boyer / Assistant Prof. (Duke)
Tim Ryan / Assoc. Prof (PSU)
Tim McGeary / Associate University Librarian (Duke)
Ed Gomes / Associate Dean of IT (Duke)
Gregg Gunnell / Director of Fossil Primates (Duke)
Seth Kaufman / CEO & Founder Whirligig Inc.

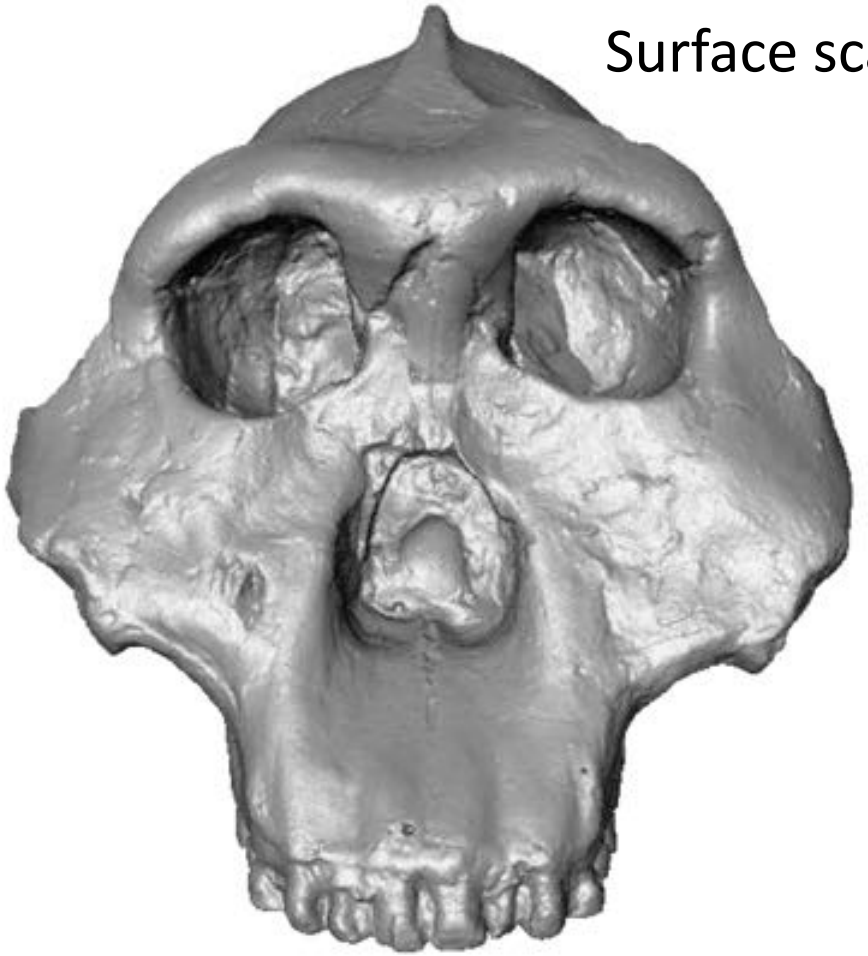
3D data representing museum specimens



microCT scans



Surface scans



www.morphosource.org

Characteristics of 3D data

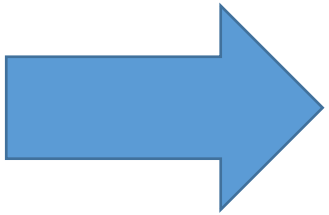
- Time consuming to generate
- Detailed metadata
- Require specialized software
- Large file sizes
- Often serve as a replacement / improvement upon the actual specimen

In a perfect world…

- All relevant data on all the world's specimens would be available at the 'click of the mouse'







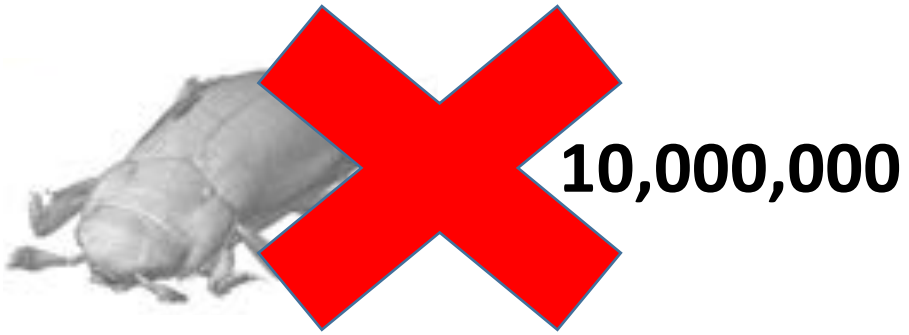

The enemy of Perfection

Blockades to ‘Digital Utopia’

- Not everyone feels that universal access to data is good
- Logistics of building, managing and maintaining such a large and diverse archive are infeasible

3D data are not like digitized specimen records

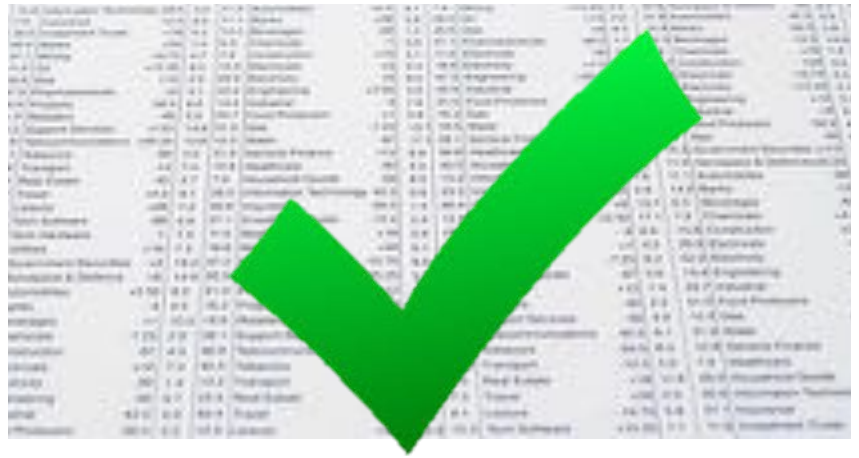
iDigBio
Integrated Digitized Biocollections

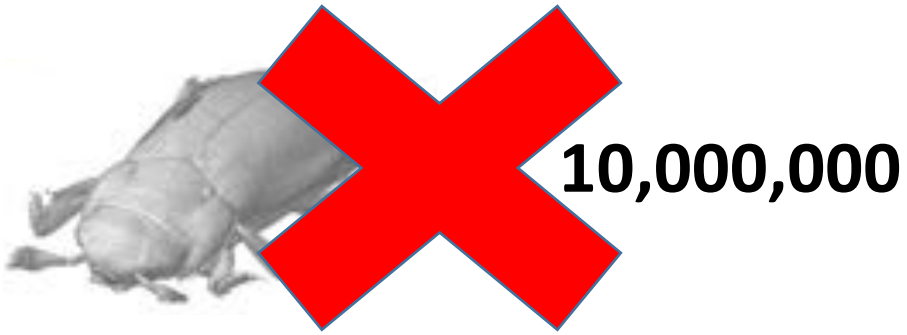


3D data are not like digitized specimen records

For 3D digitization focus…

- **NOT** on comprehensive coverage
- But on **HIGH VALUE** coverage





What are High value 3D data?

Sp

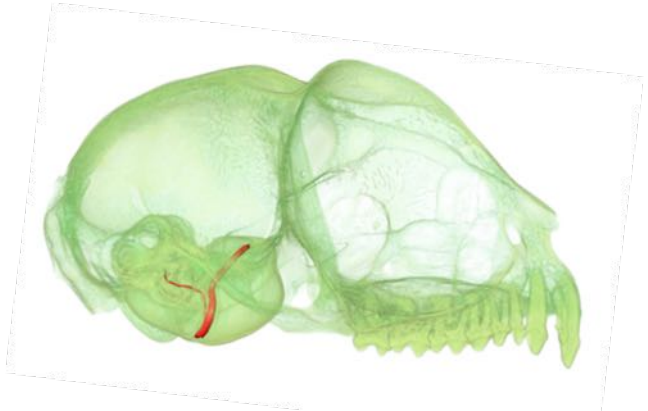
- T
- P
- R
- A
  b
- O

DEPENDS ON RESEARCH COMMUNITY

Topics

“open” questions

1. Incentives for data sharing
2. Restricting on data use
3. Long term sustainability (data formats, growth, and governance)



www.morphosource.org

Sharing

Context

- Data transparency is trending
- Compulsory in some contexts
- But still ineffective and inefficient when it comes to morphology

Sharing

Obstacles

- Researcher investment in data collection
- Museum restrictions on distribution
- Uncertainty about ownership
- Cost/finding the appropriate archive

Sharing

Potential solutions

- Compulsion (reviewers, journals, societies, granting agencies, government)
- Positive incentives

Incentivizing

Redefine currency of credit for data

- Data collectors need recognition
- Museums need recognition

Ease concerns about adherence to use restrictions

- Display copyright licenses
- Allow data owners to vet sharing requests

Realign ‘interests’

- Science benefits by **accelerating** broad access to data
- Researchers benefit by **delaying** broad access to data they collect
- Museums benefit by having collections that draw visitors

Currencies

Researchers

- Authorship on publications

Museums

- Demonstrated collection use

Gold behind the currency

Publications

- The research activities lead to important scholarly contributions

Museum visits

- The collection is a valuable one that needs to be maintained

New Currencies reflecting same values

Data Value/popularity (Egress)

- How often is a dataset viewed or accessed
- Who accesses the data
- What is it used for

Data Impact/citation (Ingress)

- Number of papers citing each dataset
- Number of papers citing grant numbers associated with each data set

Tracking datasets on MorphoSource

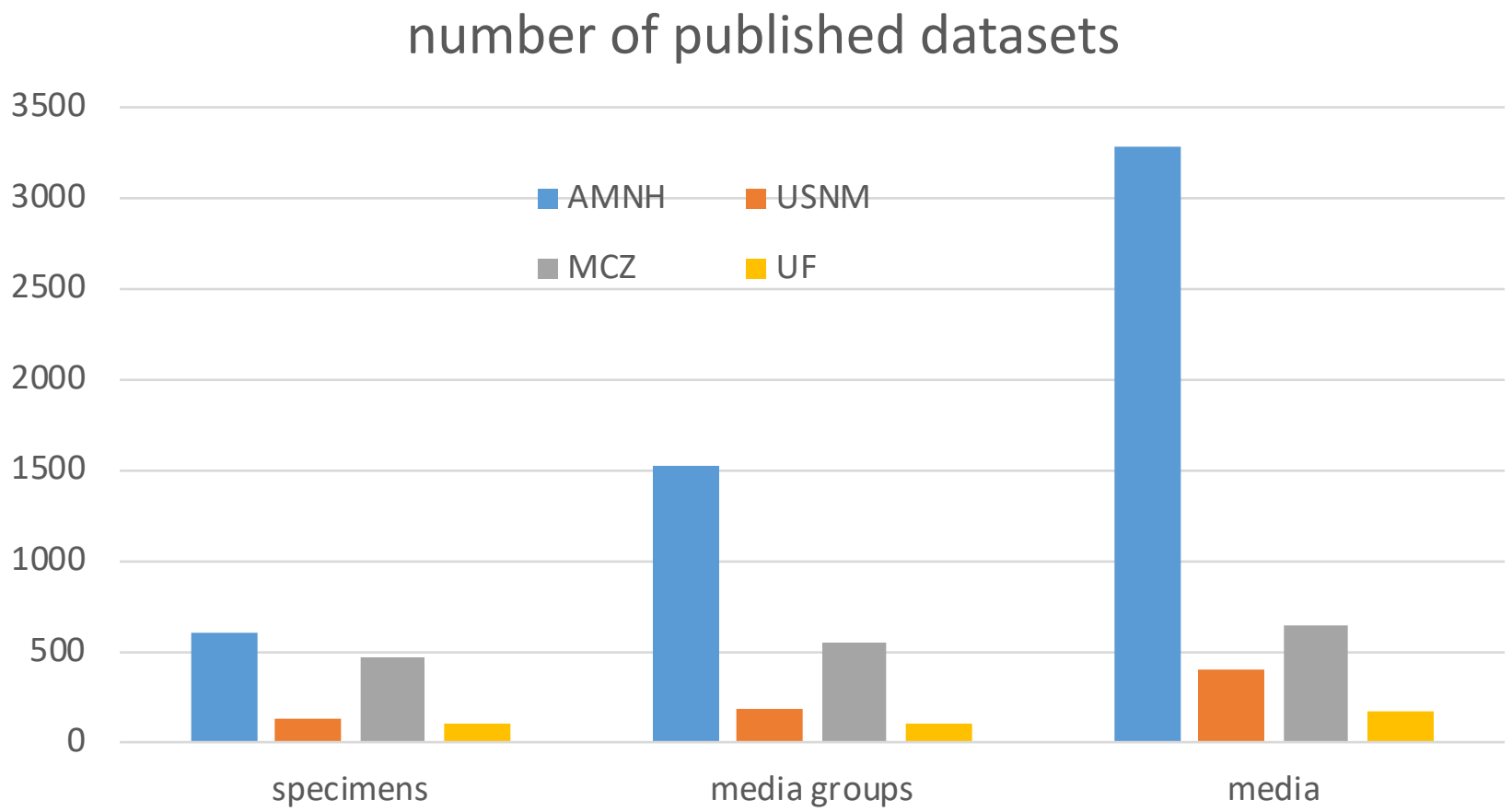
Demonstrating Data Impact - EGRESS

***Homo naledi* project**	**Harvard Primate skull project**

www.morphosource.org

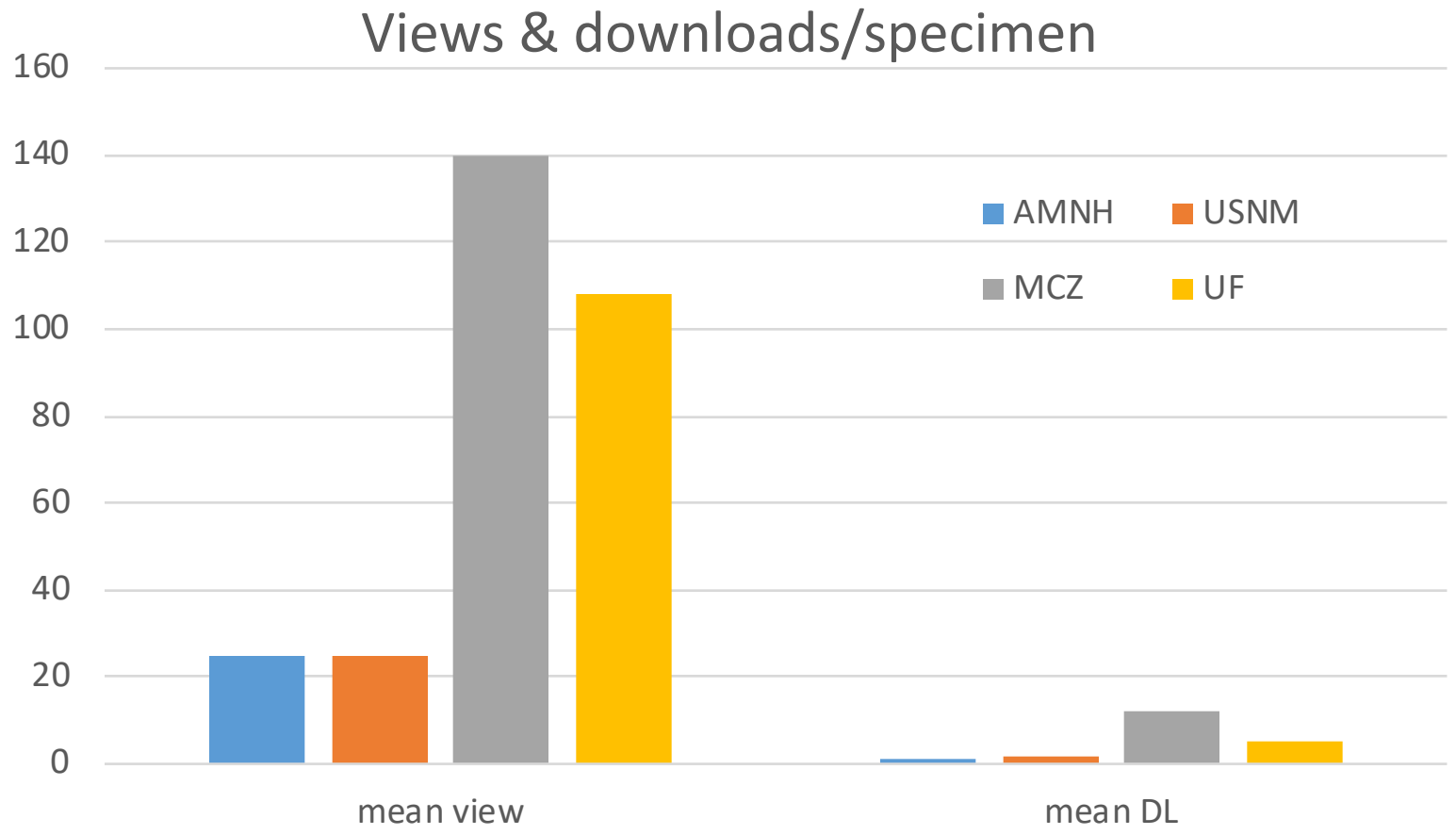
Tracking datasets on MorphoSource

Comparing Collections



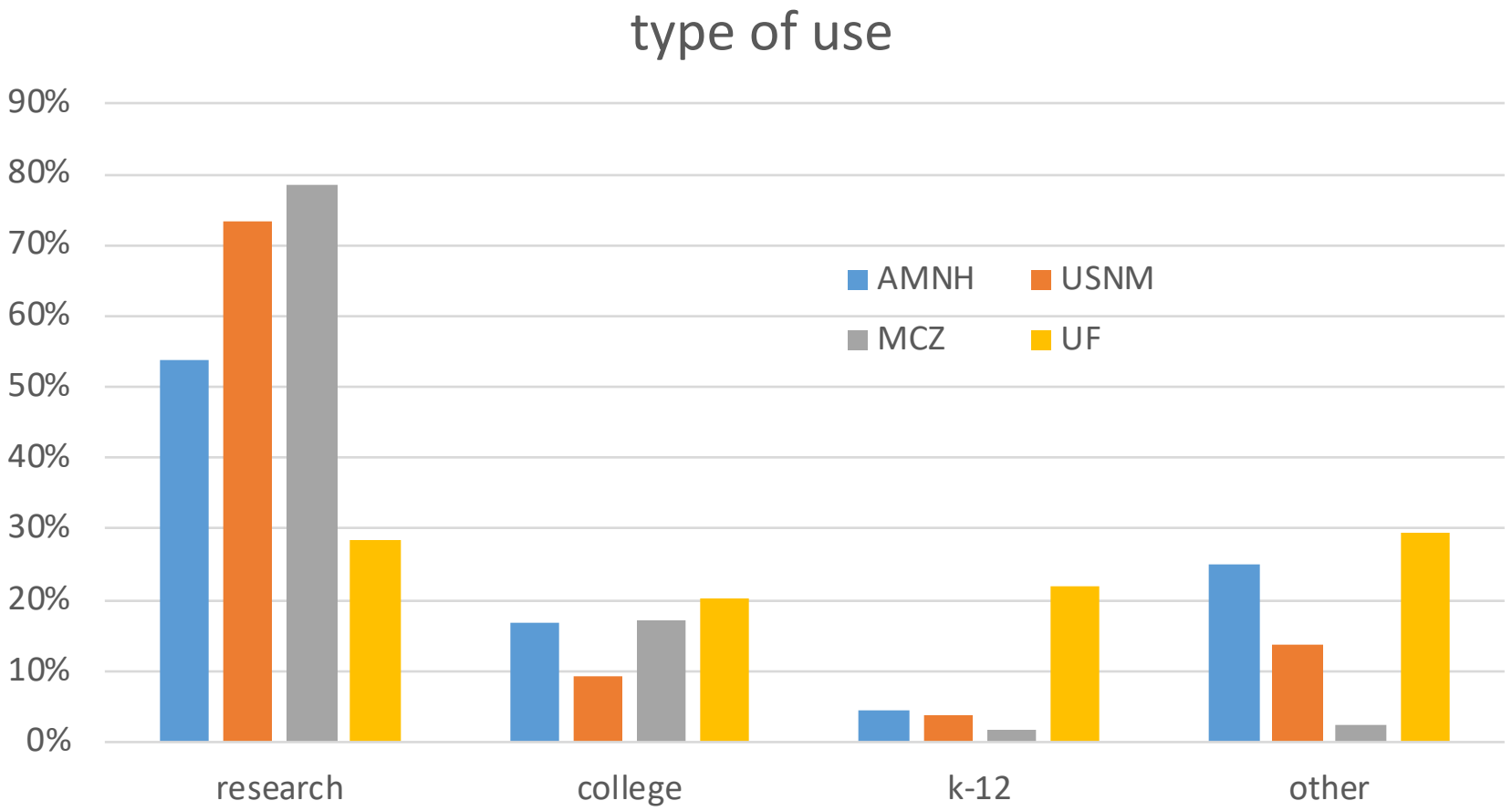
Tracking datasets on MorphoSource

Comparing Collections



Tracking datasets on MorphoSource

Comparing Collections



Tracking datasets on MorphoSource

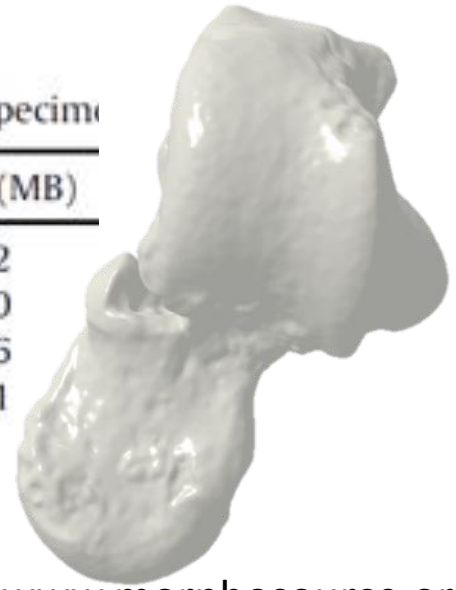
Demonstrating Data Impact - INGRESS

Tracking datasets on MorphoSource

Demonstrating Data Impact - INGRESS

Table 1
Astragali and calcanei attributed to *Anchomomys frontanyensis* and used in this study with information about each specim

Specimen	Element	Side	MoSo media	Doi	File type	File size (MB)
IPS-7712	Astragalus	Left	M6345-6065	doi:10.17602/M2/M6065	Ply, mesh file	8.02
IPS-7713	Astragalus	Left	M6346-6066	doi:10.17602/M2/M6066	Ply, mesh file	11.20
IPS-7750	Astragalus	Right	M6347-6067	doi:10.17602/M2/M6067	Ply, mesh file	10.66
IPS-7796	Astragalus	Right	M6348-6068	doi:10.17602/M2/M6068	Ply, mesh file	10.41

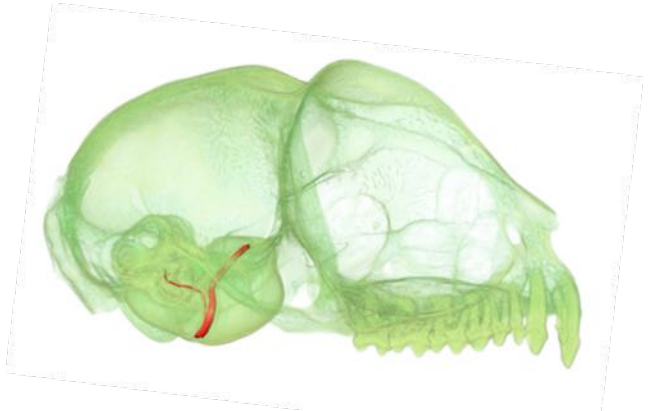


www.morphosource.org

Topics

“open” questions

1. Incentives for data sharing
2. **Restrictions on data use (copyright)**
3. Long term sustainability (data formats, growth, and governance)



www.morphosource.org

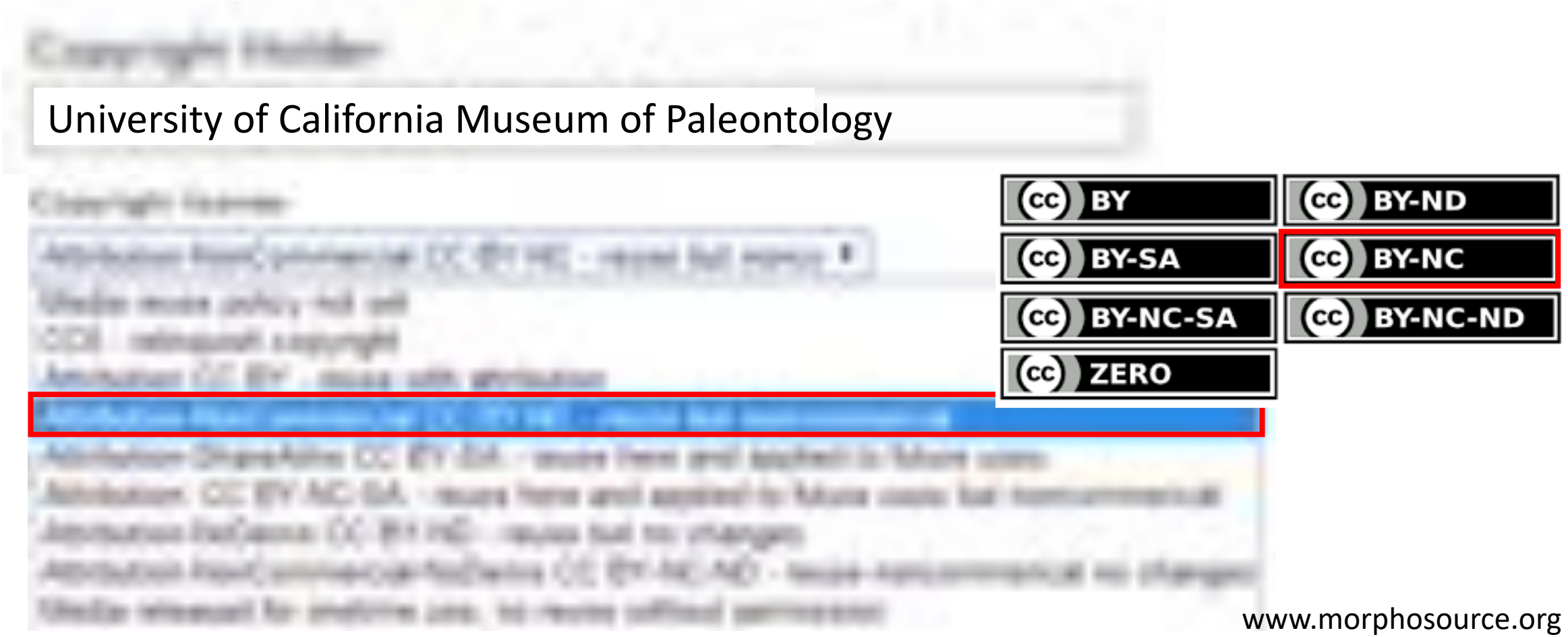
Restricting access

Ease concerns about adherence to use restrictions/attribution

- Display copyright licenses
- Track downloader identities
- Allow data owners to vet sharing requests
- What if the researcher isn't the data owner?

Copyright and ownership

Display copyright status

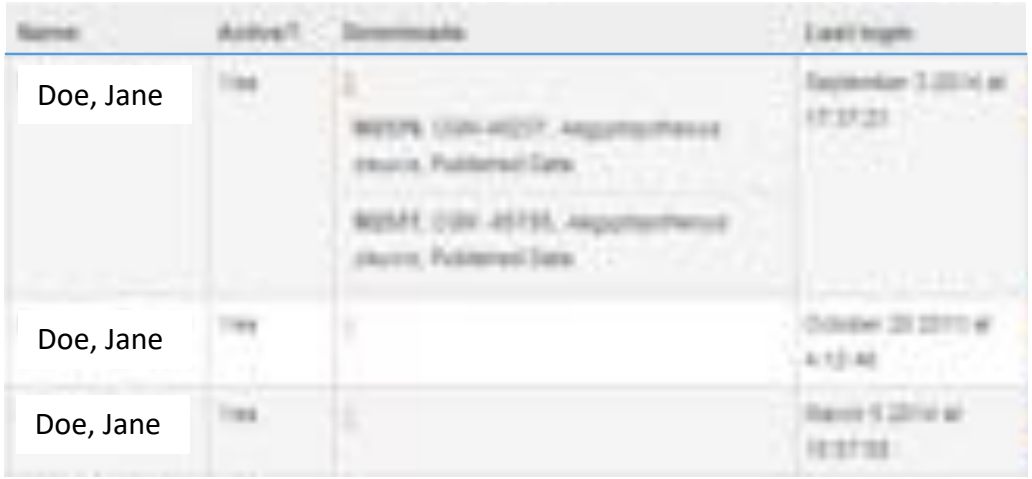


www.morphosource.org

Tracking datasets on MorphoSource

Store user identity

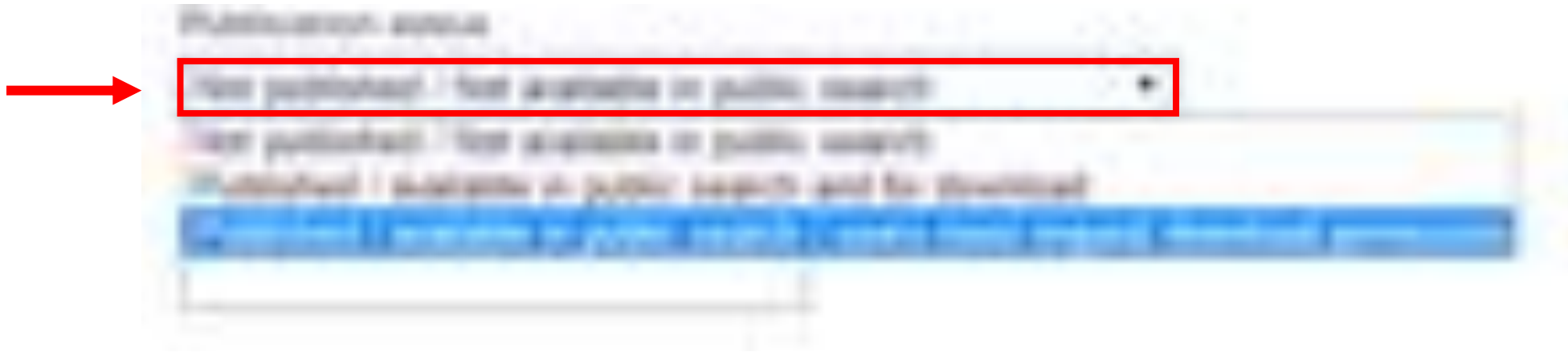
- Information available for contributors
  - Name/Institution of users downloading media file
  - Number of views, downloads, and download requests of each media file
  - Intended use (research, education, etc.)



Restricting datasets on MorphoSource

Set sharing restrictions

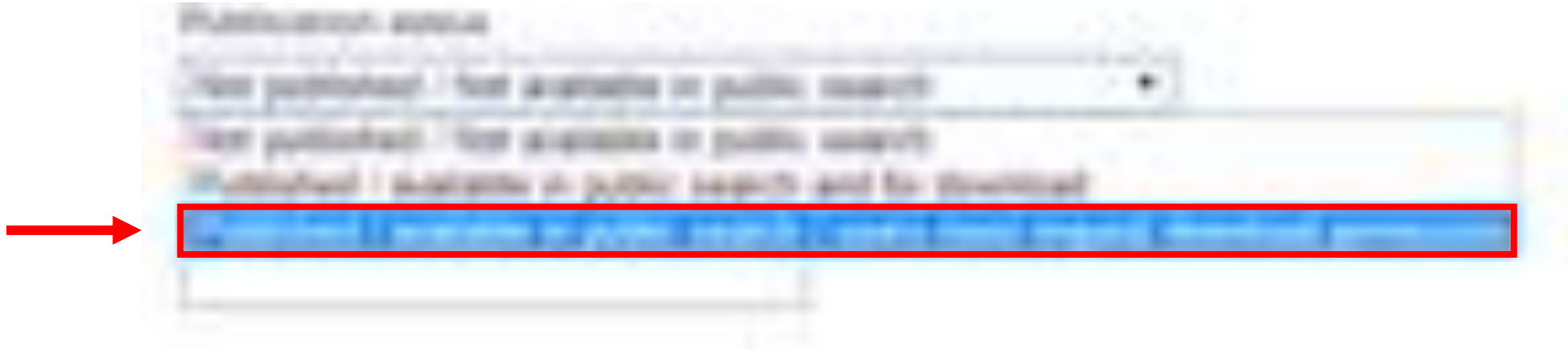
- When first uploaded, datasets are unpublished/private by default
  - Only contributor and chosen collaborators can view



Restricting datasets on MorphoSource

Set sharing restrictions

- Published – downloadable with data author permission
  - Specimen/media file returned in public search
  - Mesh files can be previewed in 3D in browsers
  - Users can send a form email request to data author for 1 time download



Sharing datasets on MorphoSource

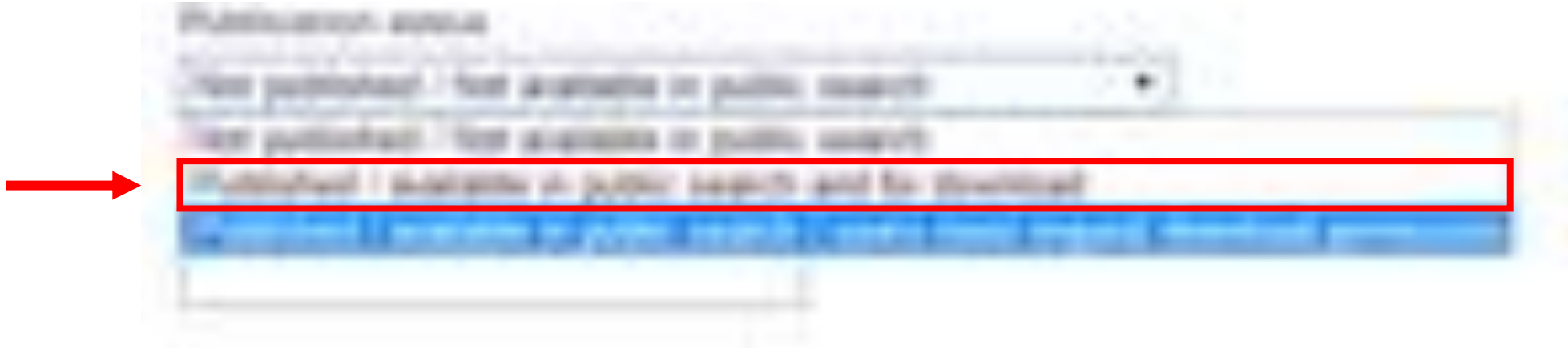
Who receives this request?

- Project manager can specify which members.
- Museum curator accounts can be added

Sharing datasets on MorphoSource

Set sharing restrictions

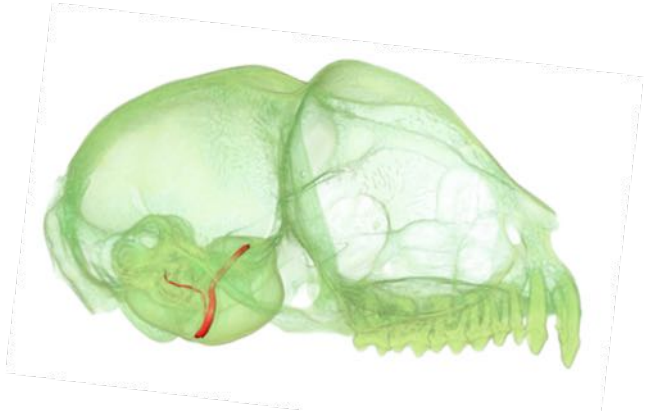
- Published – unlimited download
  - Specimen/media file returned in public search
  - Downloadable by any registered user



Outline

“open” questions

1. Incentives for data sharing
2. Restrictions on data use (copyright)
3. **Long term sustainability (data formats, growth, and governance)**

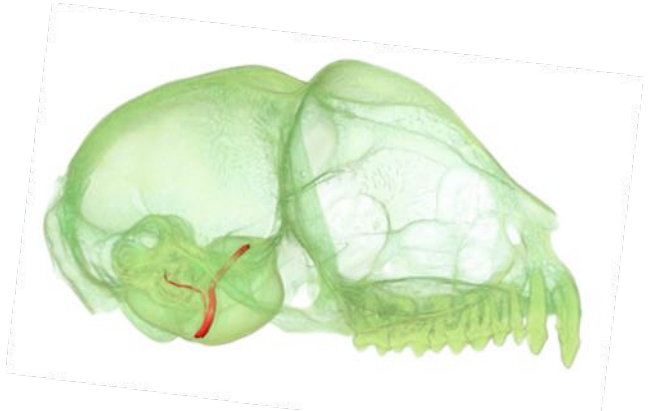


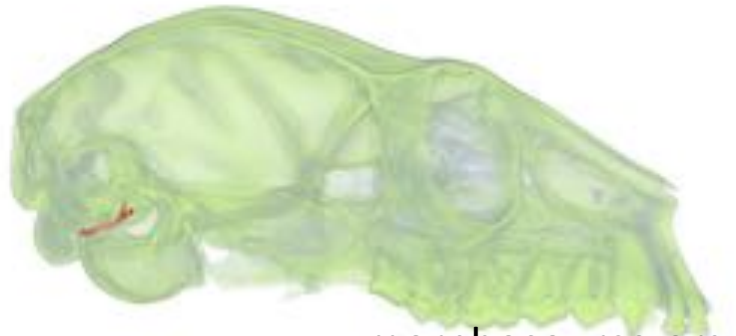
www.morphosource.org

Sustainability

Components

1. Data formats/quality control
2. Growth & Governance



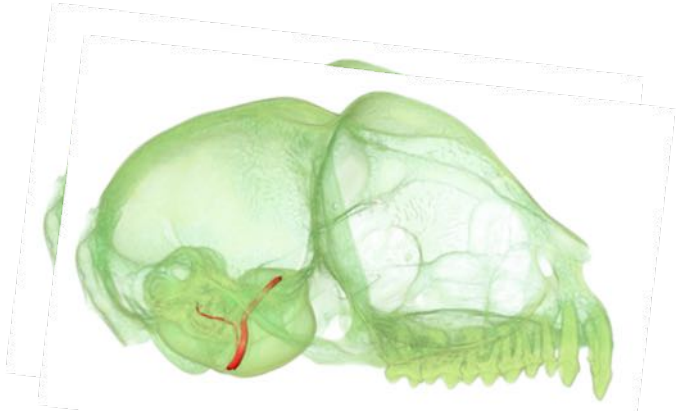


www.morphosource.org

Format Standards

Considerations

- Stability (e.g., tiff better than jpeg)
- Readability (proprietary, breadth of support, popularity)
- Efficiency (bits of data for a given quality)

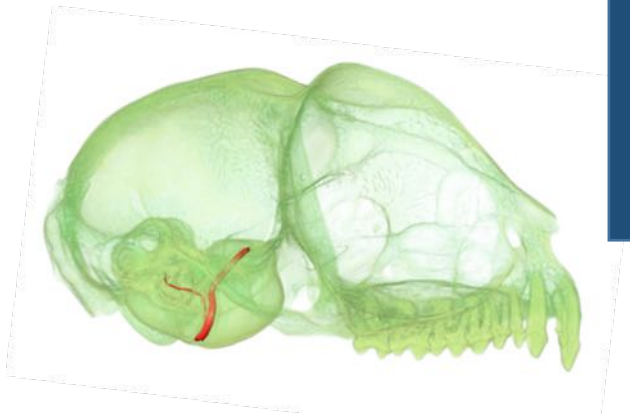


www.morphosource.org

Format Standards

Importance

- Act of specifying standards increases useability and sustainability
- Archiving initiatives ‘Archivematica’
  - Specify archive and access formats
  - Specify translation protocols between formats



archivematica

www.morphosource.org

Format Standards

Recommendations

- Surveys of MorphoSource user community

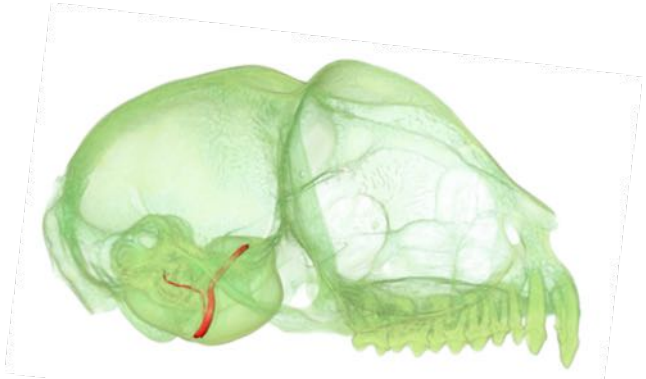
  Surface data (20% have no preference)

  - 1,700 responses: 52% stl / 23% obj / 18% ply

  Volume data (60% have no preference)

  - 790 responses: 21% dicom / 19% 16bit Tiff

- Proclamation by experts in research community



www.morphosource.org

Format Standards

Recommendations

Surface data

Stl, ply, obj, vtk

Volume data

Tiff, dicom, bmp

PROCEEDINGS B

rspb.royalsocietypublishing.org

Perspective

Cite this article: Davies TG *et al.* 2017 Open data and digital morphology. *Proc. R. Soc. B* 20170194.
http://dx.doi.org/10.1098/rspb.2017.0194

Received: 30 January 2017
Accepted: 10 March 2017

Open data and digital morphology

Thomas G. Davies[1], Imran A. Rahman[1,2], Stephan Lautenschlager[1,3], John A. Cunningham[1], Robert J. Asher[4], Paul M. Barrett[5], Karl T. Bates[6], Stefan Bengtson[7], Roger B. J. Benson[8], Doug M. Boyer[9], José Braga[10,11], Jen A. Bright[12,13], Leon P. A. M. Claessens[14], Philip G. Cox[15], Xi-Ping Dong[16], Alistair R. Evans[17], Peter L. Falkingham[18], Matt Friedman[19], Russell J. Garwood[5,20], Anjali Goswami[21], John R. Hutchinson[22], Nathan S. Jeffery[6], Zerina Johanson[5], Renaud Lebrun[23], Carlos Martínez-Pérez[1,24], Jesús Marugán-Lobón[25], Paul M. O'Higgins[15], Brian Metscher[26], Maëva Orliac[23], Timothy B. Rowe[27], Martin Rücklin[1,28], Marcelo R. Sánchez-Villagra[29], Neil H. Shubin[30], Selena Y. Smith[19], J. Matthias Starck[31], Chris Stringer[5], Adam P. Summers[32], Mark D. Sutton[33], Stig A. Walsh[34], Vera Weisbecker[35], Lawrence M. Witmer[36], Stephen Wroe[37], Zongjun Yin[1,38], Emily J. Rayfield[1] and Philip C. J. Donoghue[1]

Quality Control

Documentation

- Quality is relative
- Different research methods differ
- Lack of metadata

PROCEEDINGS B

rspb.royalsocietypublishing.org

Perspective

Cite this article: Davies TG *et al.* 2017 Open data and digital morphology. *Proc. R. Soc. B* 20170194.
http://dx.doi.org/10.1098/rspb.2017.0194

Received: 30 January 2017
Accepted: 10 March 2017

Open data and digital morphology

Thomas G. Davies[1], Imran A. Rahman[1,2], Stephan Lautenschlager[1,3], John A. Cunningham[1], Robert J. Asher[4], Paul M. Barrett[5], Karl T. Bates[6], Stefan Bengtson[7], Roger B. J. Benson[8], Doug M. Boyer[9], José Braga[10,11], Jen A. Bright[12,13], Leon P. A. M. Claessens[14], Philip G. Cox[15], Xi-Ping Dong[16], Alistair R. Evans[17], Peter L. Falkingham[18], Matt Friedman[19], Russell J. Garwood[5,20], Anjali Goswami[21], John R. Hutchinson[22], Nathan S. Jeffery[6], Zerina Johanson[5], Renaud Lebrun[23], Carlos Martínez-Pérez[1,24], Jesús Marugán-Lobón[25], Paul M. O'Higgins[15], Brian Metscher[26], Maëva Orliac[23], Timothy B. Rowe[27], Martin Rücklin[1,28], Marcelo R. Sánchez-Villagra[29], Neil H. Shubin[30], Selena Y. Smith[19], J. Matthias Starck[31], Chris Stringer[5], Adam P. Summers[32], Mark D. Sutton[33], Stig A. Walsh[34], Vera Weisbecker[35], Lawrence M. Witmer[36], Stephen Wroe[37], Zongjun Yin[1,38], Emily J. Rayfield[1] and Philip C. J. Donoghue[1]

Growth

- CT
- Sy
- M
- Ho
- Fo
  1
  2
  3

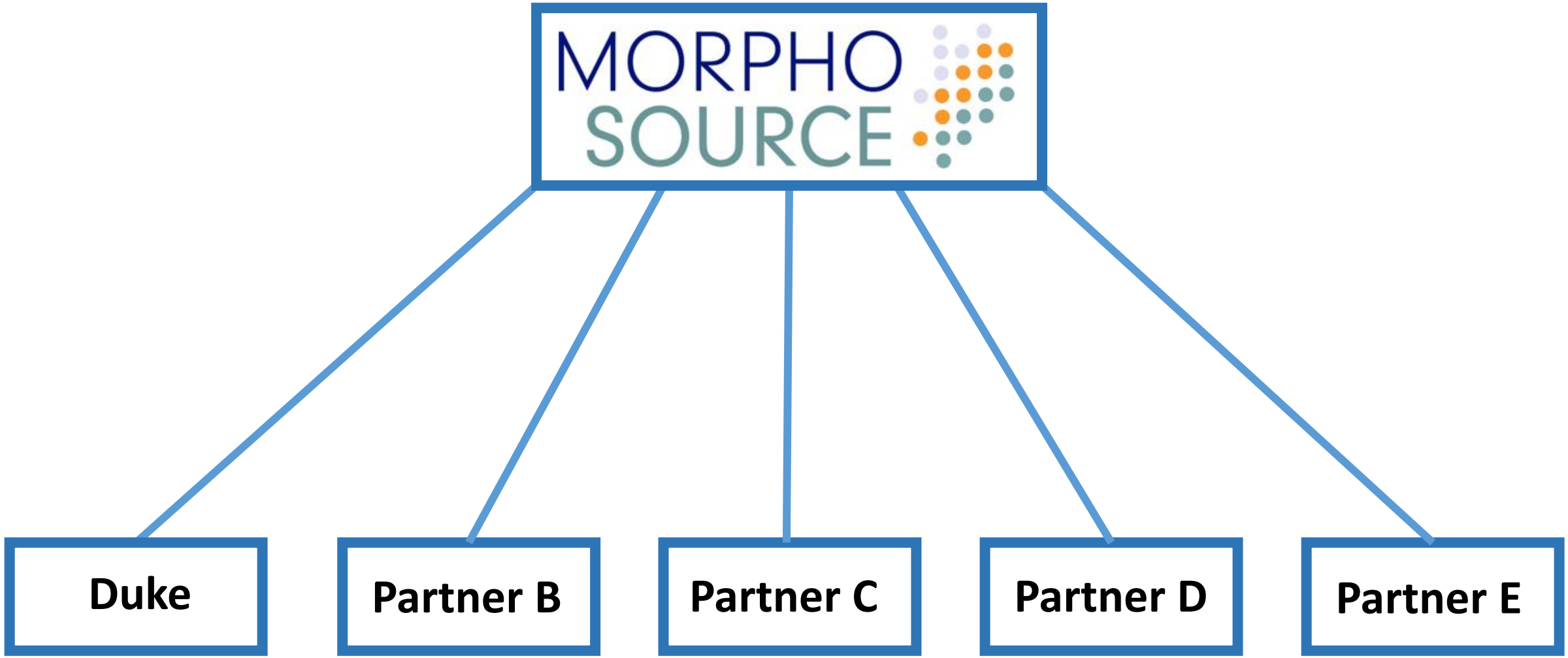
Growth

Power in numbers…. Many hands make light work

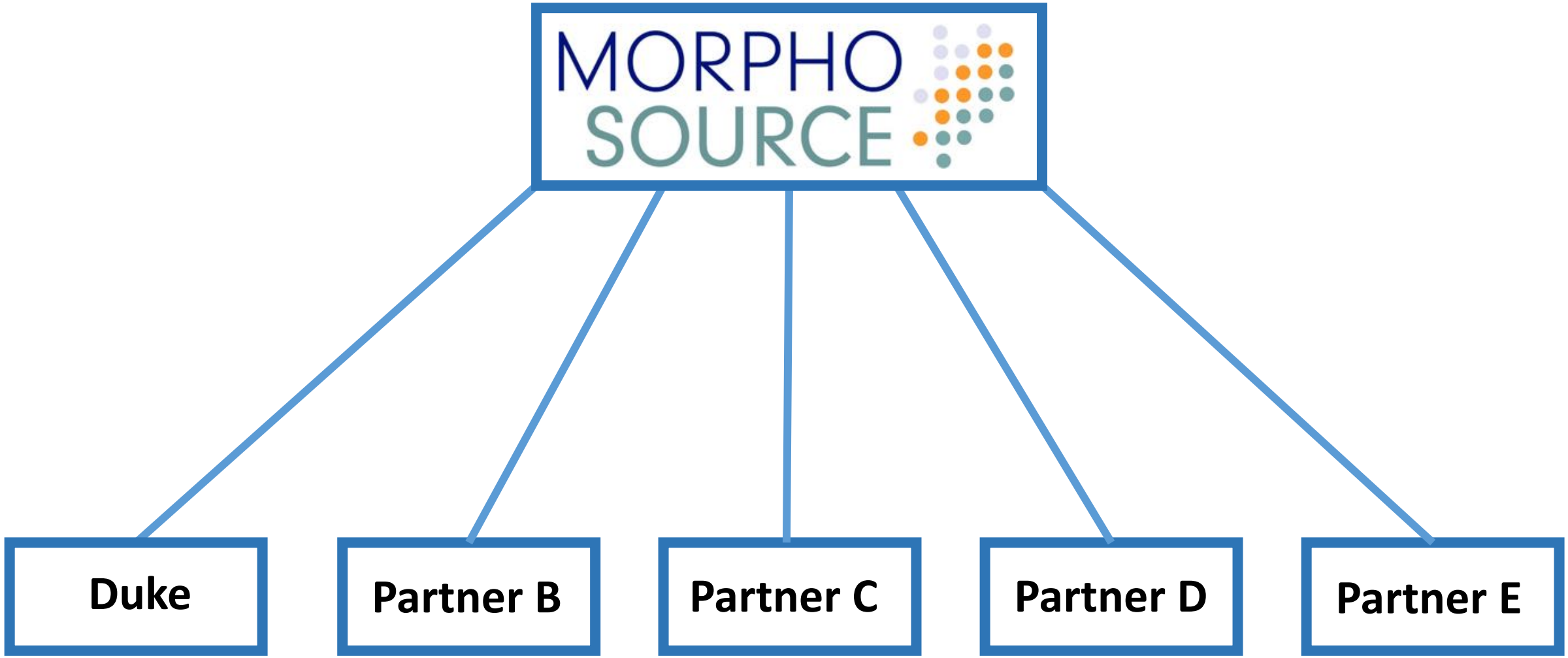
Benefits of using a consortium to grow

- Commitment from consortium members can be finite
- Redundancy of data across wide geographic areas
- Many partners have the infrastructure to take over management

Growth



How?

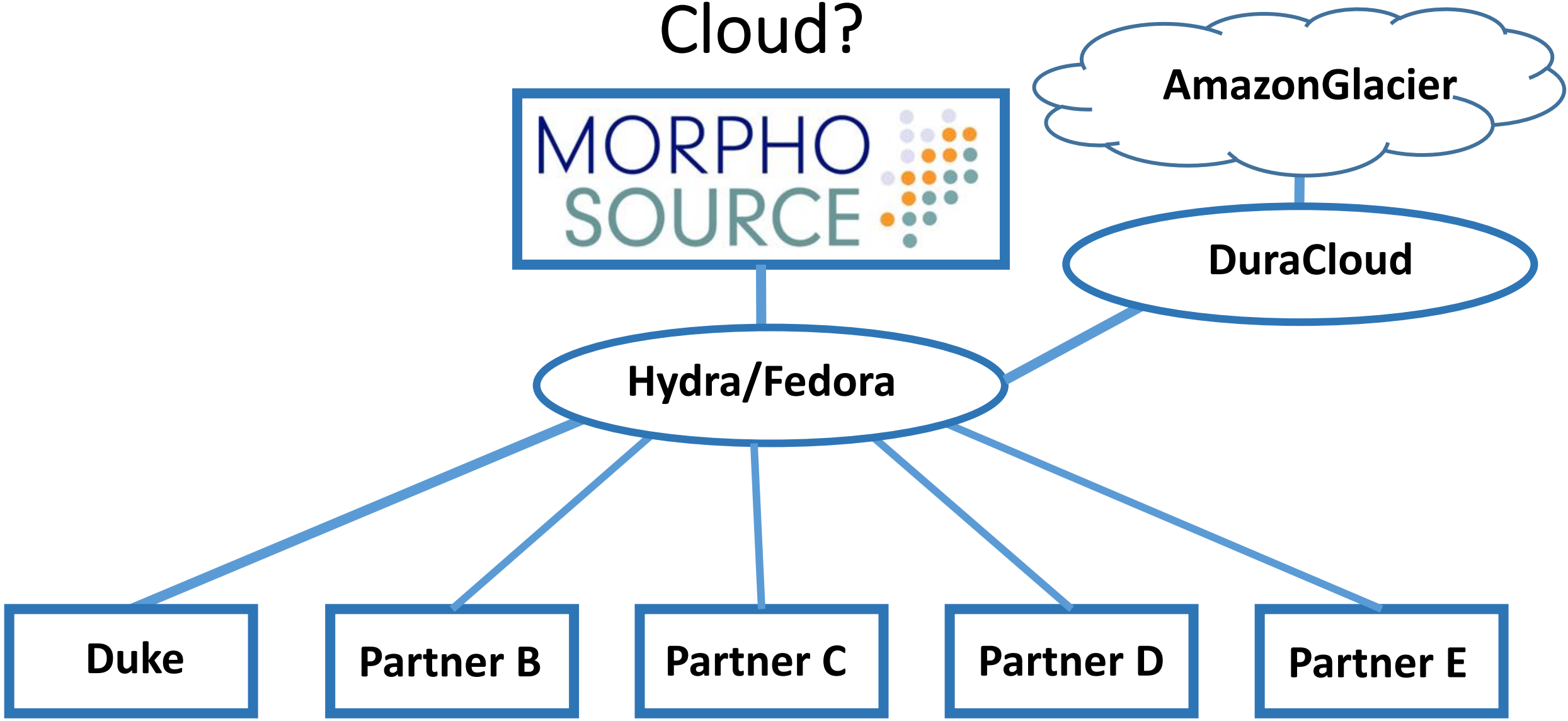


How?

Leverage state-of-the-art open access digital repository platforms

Duke | **Partner B** | **Partner C** | **Partner D** | **Partner E**

Cloud?


MORPHO SOURCE
AmazonGlacier
DuraCloud
Hydra/Fedora
Duke
Partner B
Partner C
Partner D
Partner E

The open source parts and their functions

Fedora (Fedorarepository.org)

- Digital Asset Management platform
- Allows integration of multiple data nodes

Hydra (projecthydra.org)

- Provides the database structure for interacting with Fedora

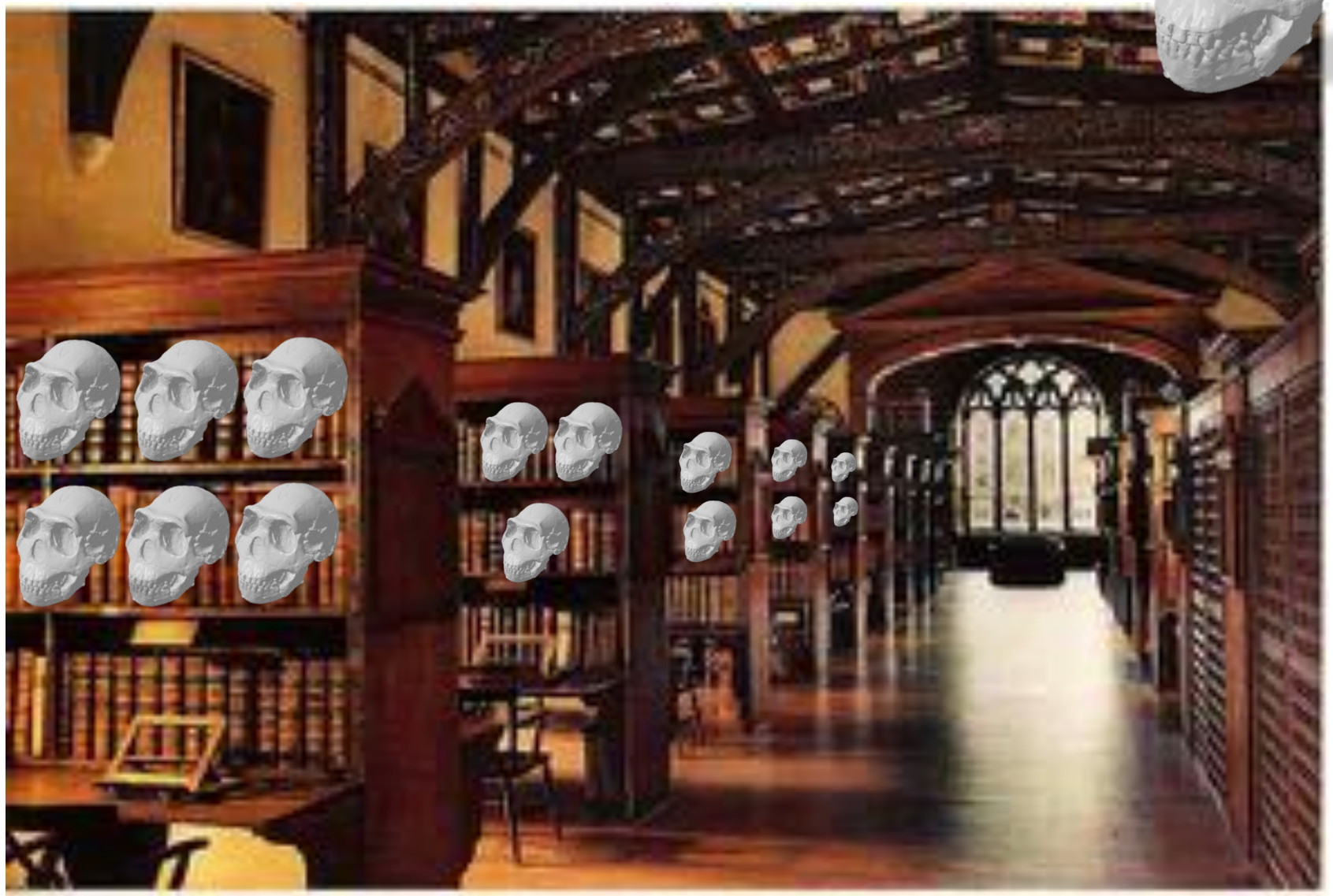
DuraCloud (duracloud.org)

- Provides bitrot prevention
- Manages full redundant cloud copy through Amazon Glacier
- Provides integration with Archivematica

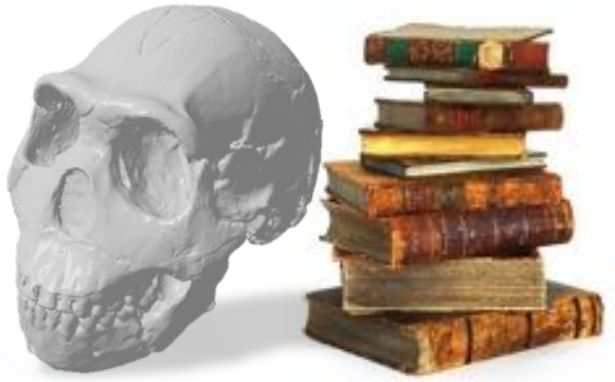
‘the consortium’

- Manages the cost of a growing community data archive
- Provides geographic robustness
- Provides options for shifting governance

Library Partnership



Summary



3D data in Paleontology

1. Incentivizing data sharing is key
2. Successful databases must allow sharing restrictions
3. Formats and standards should be clearly defined
4. Integrated community governance and support should be sought for longterm sustainability

Acknowledgments

For invitation to speak
- Pat Holroyd, Talia Karim, Gil Nelson

For support & funding of MorphoSource Development
- Duke University Trinity College of Arts & Sciences (major funder so far)
- Duke Shared Materials Instrumentation Facility
- Duke Biology IT Center
- Ed Gomes & Trinity Technology Services

For discussion leading to development of concepts
- Jukka Jernvall, Alistair Evans, & Gudrun Evans

For work loading specimen media
- Technicians & students: Mercedes Zapata-Garcia, Shane Daly, Sunghoon Liu, Ksenia Sokolova, Anne Driscoll, Kevin Vo, Annie Lott, Callie Crawford, and many more.