Enumerating and Addressing AI-Readiness Challenges from the Biology Guided Neural Networks HDR Project

Joel Pepper[1]

Additional Authors of Papers I'm Covering: Jane Greenberg[1], David Breen[1], Andrew Senin[1], Kevin Karnani[1], Ajani Levere[1], Jeremy Leipzig[1], Dom Jebbia[2], Yasin Bakiş[3], Xiaojun Wang[3], Henry Bart[3]

[1]Drexel University Metadata Research Center

[2]Carnegie Melon University Libraries, [3]Tulane University Biodiversity Research Institute

April 15, 2024

DREXEL UNIVERSITY Metadata Research Center College of Computing & Informatics

NSF-OAC# 1940233

Outline

Biology Guided Neural Networks Background

- **BGNN Goal**: To develop novel machine learning tools to analyze digitized specimen images, with an eye towards using these to answer scientific questions on a large population scale.
- **Data Sources**: The group ended up focusing on repositories of fish specimen images due to their availability (University of Illinois, Ohio State, Field Museum, University of Wisconsin, iDigBio meta-collection)

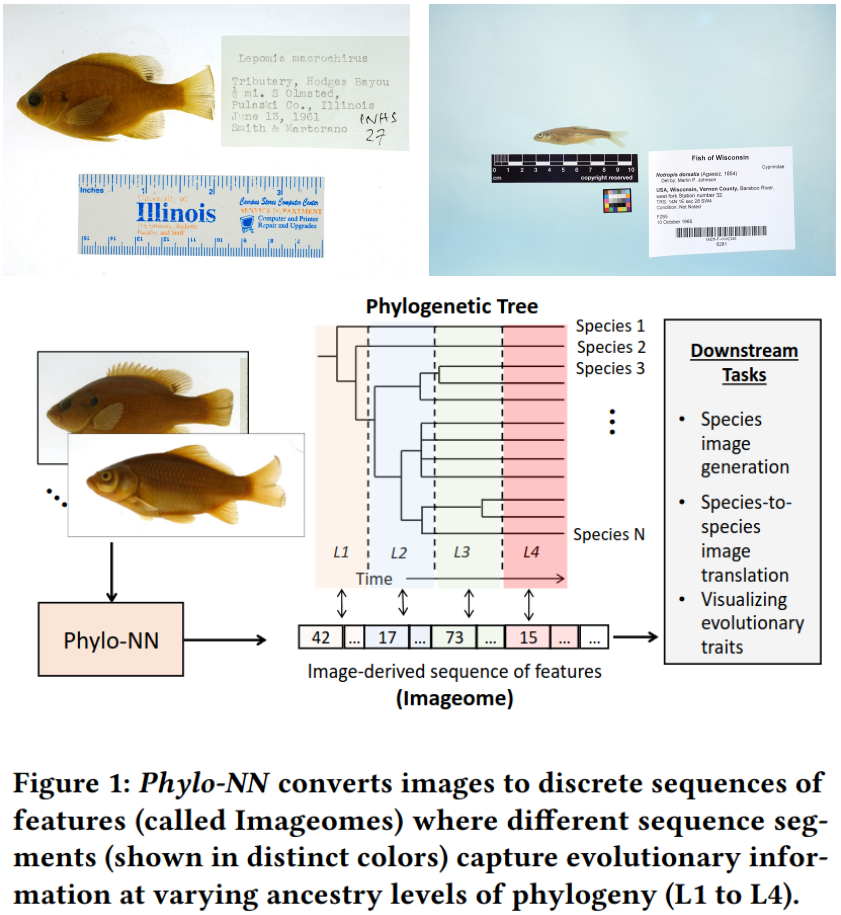


Figure 1: *Phylo-NN* converts images to discrete sequences of features (called Imageomes) where different sequence segments (shown in distinct colors) capture evolutionary information at varying ancestry levels of phylogeny (L1 to L4).

Figure: (Top) Example specimen images. (Bottom) Model developed by Virginia Tech [3].

Early Metadata Challenges

- Metadata quality issues beset the project early on, including: non-fish specimen labeled as fish, missing metadata, insufficient metadata to exclude inadmissible images (partial, damaged or occluded fish, poor image contrast), and difficulties with mass downloading and categorizing of images.
- To an extent, the group had expected iDigBio would be a one stop shop for images, but they ended up having to go to a variety of primary repositories then collate the disparate collections.
- **Big Takeaway:** Major issues/slowdowns stemming from metadata quality were encountered that were not budgeted for either temporally or monetarily in the original funding proposal, which hampered all teams involved in the BGNN project.

Image Quality Metadata Paper

- The Drexel team, in charge of managing and working on the metadata aspects of the project, set out to specify what issues needed fixed to ameliorate the aforementioned problems.
- The first paper from our group, in collaboration with the biology team at Tulane, developed a set of 22 properties that ideally need to be known for each image to determine their usability for ML applications.
- These properties were then manually labeled for a subset of the Illinois Natural History Survey fish image collection.

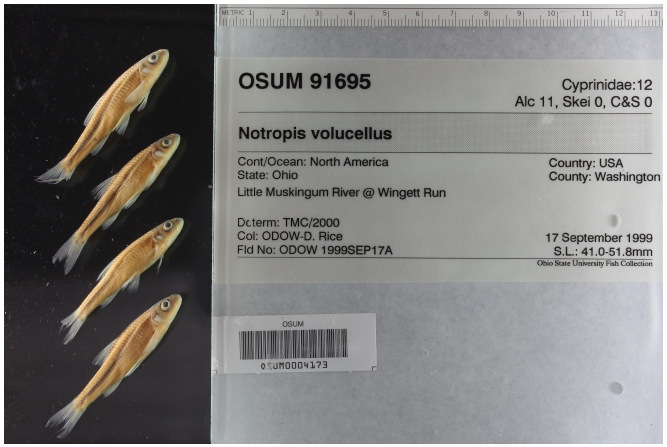
Supplemental table 1. Quality Metadata Fields

Metadata: Quality property	Data type	Description
1. if_fish	boolean	Whether or not fish are in the image.
2. fish_number	integer	Number of fish in the image.
3. if_ruler	boolean	Whether or not a ruler in the image.
4. if_label	boolean	Whether or not a label in the image.
5. if_label_name_correct	boolean	Whether or not the image name matches the number on the label.
6. if_label_catalog_number_correct	boolean	Whether or not the catalog number is correct.
7. label_detailed	categorical	Whether or not the label is detailed and has full information.
8. if_colorbar	boolean	Whether or not a colorbar is in the image.

Image Metadata Entry

Primary Metadata Problems to Address

- A lot of work was being done to download, sort and filter images from various repositories.
- Our goal at Drexel became to automate as much of the image quality and property assessment as possible, and to develop generalizable tools for accomplishing this.
- Important properties included: Specimen count, orientation, size, bounding box, and image foreground/background contrast level.



Automatic Metadata Generation Paper

- We trained an object detection tool (Detectron) to find fish, fish eyes, and the ruler within specimen images.
- We used this information to determine a number of the properties in the table on the right.
- We refined the bounding box and outline results from Detectron to generate tight masks over the specimen.

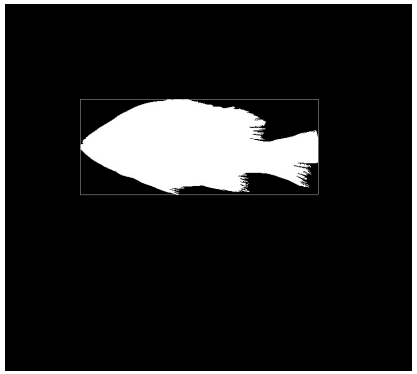
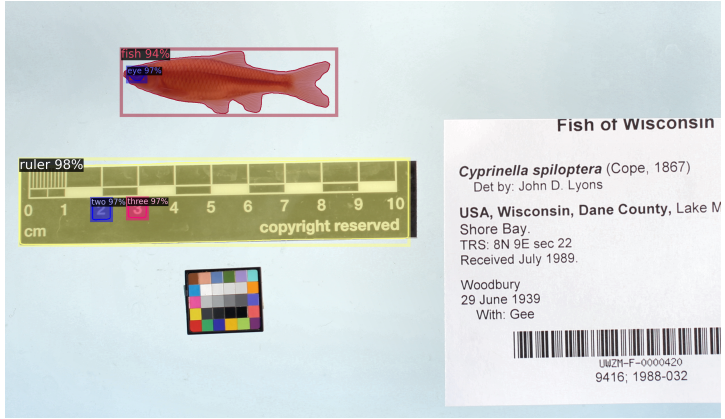


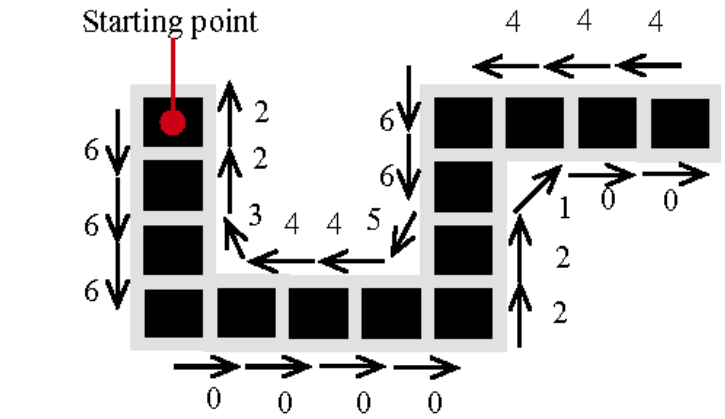
TABLE II
METADATA PROPERTIES (* INDICATES HIGHER ORDER DERIVED PROPERTIES)

Property	Association	Type	Explanation
has_fish	Overall Image	Boolean	Whether a fish was found in the image.
fish_count	Overall Image	Integer	The quantity of fish present.
has_ruler	Overall Image	Boolean	Whether a ruler was found in the image.
ruler_bbox	Overall Image	4 Tuple	The bounding box of the ruler (if found).
scale*	Overall Image	Float	The scale of the image in $\frac{\text{pixels}}{\text{cm}}$.
bbox	Per Fish	4 Tuple	The top left and bottom right coordinates of the bounding box for a fish.
background.mean	Per Fish	Float	The mean intensity of the background within a given fish's bounding box.
background.std	Per Fish	Float	The standard deviation of the background within a given fish's bounding box.
foreground.mean	Per Fish	Float	The mean intensity of the foreground within a given fish's bounding box.
foreground.std	Per Fish	Float	The standard deviation of the foreground within a given fish's bounding box.
contrast*	Per Fish	Float	The contrast between foreground and background intensities within a given fish's bounding box.
centroid	Per Fish	4 Tuple	The centroid of a given fish's bitmask.
primary_axis*	Per Fish	2D Vector	The unit length primary axis (eigenvector) for the bitmask of a given fish.
clock_value*	Per Fish	Integer	Fish's primary axis converted into an integer "clock value" between 1 and 12.
length*	Per Fish	Float	The length of a fish in centimeters.
mask	Per Fish	2D Matrix	The bitmask of a fish in 0's and 1's.
pixel_analysis_failed	Per Fish	Boolean	Whether the pixel analysis process failed for a given fish. If true, detectron's mask and bounding box were used for metadata generation.
score	Per Fish	Float	The percent confidence score output by detectron for a given fish.
has_eye	Per Fish	Boolean	Whether an eye was found for a given fish.
eye_center	Per Fish	2 Tuple	The centroid of a fish's eye.
side*	Per Fish	String	The side (i.e. 'left' or 'right') of the fish that is facing the camera (dependent on finding its eye).

Extension Paper

- In a followup paper, we extended our software beyond the original image repository we started with.
- We also improved bounding box and mask generation accuracy through the use of foreground/background contrast enhancement.
- We also added a number of metadata properties to the automatic generation pipeline, in particular a chain encoding of the specimen outlines.

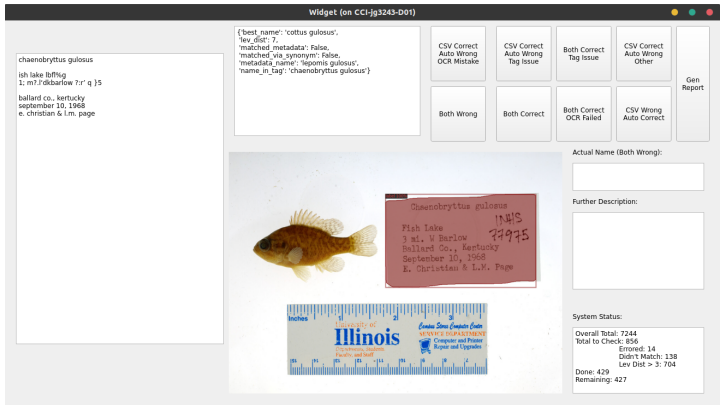




Chain Code = {66600002210044466544322}

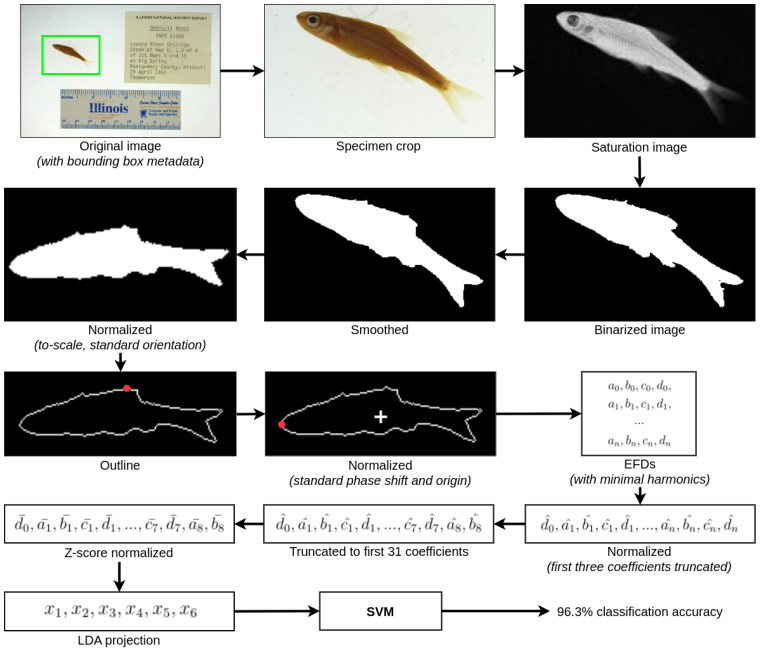
Tag Analysis Paper

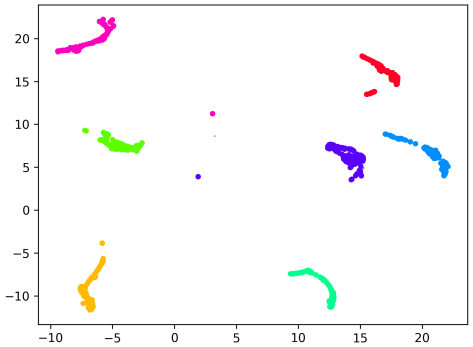
- We developed a tag analysis and accompanying GUI tool to verify the species name provided in repository metadata for specimen images.
- The name in the tag was correct essentially 100% of the time, but occasionally the metadata was incorrect, posing a big problem when training the ML tools.
- We located the tag with Detectron, then used the optical character recognition tool Tesseract to extract the tag text and analyzed it to find the scientific name and compare it to the metadata.



Specimen Outlining Paper

- Pursuant to the ultimate goal of answering scientific questions about organism populations, we began working to augment and analyze the specimen outlines we'd previously been extracting.
- By converting outline chain encoding to Fourier descriptors, we were able to develop an outline representation that through support vector machine classification or even clustering alone can be used to accurately determine specimen species.





Conclusions

- An inordinate amount of time and resources were dedicated to dealing with metadata quality issues during the BGNN project.
- Repositories of the nature utilized in this project need to give greater consideration to AI readiness and how their collections will be used in modern data science in order to facilitate more rapid scientific discovery.
- We worked to develop automated tools to ease and facilitate generation and validation of important metadata, but more work is needed to ensure AI readiness is given appropriate priority in archival science and related fields going forward.

DREXEL UNIVERSITY
Metadata Research Center
College of Computing & Informatics

NSF-OAC# 1940233

Bakiş, Y.; Altıntaş, B.; Wang, X.; Maruf, M.; Karpatne, A.; Bart, H.: Extracting landmark and trait information from segmented digital specimen images generated by artificial neural networks.
Biodiversity Information Science and Standards, 6, e94955, 2022.
http://doi.org/10.3897/biss.6.94955.

Breen, D.E.; Senin, A.; Levere, A.; Pepper, J.; Greenberg, J.: Specimen Outlining: A Computational Archival Science Approach.
In 2023 IEEE International Conference on Big Data (BigData), 2004–2009, 2023.
http://doi.org/10.1109/BigData59044.2023.10386985.

Elhamod, M.; Khurana, M.; Manogaran, H.B.; Uyeda, J.C.; Balk, M.A.; Dahdul, W.; Bakis, Y.; Bart, H.L.; Mabee, P.M.; Lapp, H.; Balhoff, J.P.; Charpentier, C.; Carlyn, D.; Chao, W.L.; Stewart, C.V.; Rubenstein, D.I.; Berger-Wolf, T.; Karpatne, A.: Discovering novel biological traits from images using phylogeny-guided neural networks.
In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '23, 3966–3978. Association for Computing Machinery, New York, NY, USA, 2023.
ISBN 9798400701030.
http://doi.org/10.1145/3580305.3599808.

Karnani, K.; Pepper, J.; Bakiş, Y.; Wang, X.; Bart Jr., H.; Breen, D.E.; Greenberg, J.: Computational metadata generation methods for biological specimen image collections.
International Journal on Digital Libraries, 2022.
ISSN 1432-1300.
http://doi.org/10.1007/s00799-022-00342-1.

Leipzig, J.; Bakis, Y.; Wang, X.; Elhamod, M.; Diamond, K.; Dahdul, W.; Karpatne, A.; Maga, M.; Mabee, P.; Bart, H.L.; Greenberg, J.: Biodiversity image quality metadata augments convolutional neural network classification of fish species.
In E. Garoufallou; M.A. Ovalle-Perandones, eds., Metadata and Semantic Research, 3–12. Springer International Publishing, Cham, 2021.
ISBN 978-3-030-71903-6.
http://doi.org/10.1007/978-3-030-71903-6_1.

Pepper, J.; Greenberg, J.; Bakiş, Y.; Wang, X.; Bart, H.; Breen, D.: Automatic Metadata Generation for Fish Specimen Image Collections.
In 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 31–40, 2021.
http://doi.org/10.1109/JCDL52503.2021.00015.

Pepper, J.; Senin, A.; Jebbia, D.; Breen, D.; Greenberg, J.: Metadata Verification: A Workflow for Computational Archival Science.

In 2022 IEEE International Conference on Big Data (Big Data), 2565–2571, 2022. http://doi.org/10.1109/BigData55660.2022.10020340.