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

Joel Pepper\textsuperscript{1}

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

\textsuperscript{1}Drexel University Metadata Research Center
\textsuperscript{2}Carnegie Melon University Libraries, \textsuperscript{3}Tulane University Biodiversity Research Institute

April 15, 2024
Outline

1. Biology Guided Neural Networks Background

2. Year 1 (Before I Joined)
   - *Biodiversity Image Quality Metadata Augments Convolutional Neural Network Classification of Fish Species* [5]

3. Years 2-5
   - *Automatic Metadata Generation for Fish Specimen Image Collections* [6]
   - *Computational metadata generation methods for biological specimen image collections* [4]

4. Conclusions
**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).
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.
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.
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.
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.
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\}
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.
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.
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.
Biodiversity Information Science and Standards, 6, e94955, 2022.
http://doi.org/10.3897/biss.6.94955.

http://doi.org/10.1109/BigData59044.2023.10386985.

ISBN 9798400701030.
ISSN 1432-1300. 

http://doi.org/10.1007/978-3-030-71903-6_1.

http://doi.org/10.1109/JCDL52503.2021.00015.