On Image Quality Metadata, FAIR in ML, AI-Readiness and Reproducibility: Fish-AIR example

Yasin Bakış[‡], Xiaojun Wang[‡], Bahadır Altıntaş[‡,§], Dom Jebbia[|], Henry L Bart Jr.[‡]

‡ Tulane University, New Orleans, United States of America
§ Abant Izzet Baysal Universitesi, Bolu, Turkiye
| Carnegie Mellon University, Pittsburgh, United States of America

Corresponding author: Yasin Bakış (yasinbakis@gmail.com)

Abstract

A new science discipline has emerged within the last decade at the intersection of informatics, computer science and biology: *Imageomics*. Like most other -omics fields, Imageomics also uses emerging technologies to analyze biological data but from the images. One of the most applied data analysis methods for image datasets is Machine Learning (ML). In 2019, we started working on a United States National Science Foundation (NSF) funded project, known as Biology Guided Neural Networks (BGNN) with the purpose of extracting information about biology by using neural networks and biological guidance such as species descriptions, identifications, phylogenetic trees and morphological annotations (Bart et al. 2021). Even though the variety and abundance of biological data is satisfactory for some ML analysis and the data are openly accessible, researchers still spend up to 80% of their time preparing data into a usable, AI-ready format, leaving only 20% for exploration and modeling (Long and Romanoff 2023). For this reason, we have built a dataset composed of digitized fish specimens, taken either directly from collections or from specialized repositories. The range of digital representations we cover is broad and growing, from photographs and radiographs, to CT scans, and even illustrations. We have added new groups of vocabularies to the dataset management system including image quality metadata, extended image metadata and batch metadata. With the image quality metadata and extended image metadata, we aimed to extract information from the digital objects that can possibly help ML scientists in their research with filtering, image processing and object recognition routines. Image quality metadata provides information about objects contained in the image, features and condition of the specimen, and some basic visual properties of the image, while extended image metadata provides information about technical properties of the digital file and the digital multimedia object (Bakış et al. 2021, Karnani et al. 2022, Leipzig et al. 2021, Pepper et al. 2021, Wang et al. 2021) (see details on Fish-AIR vocabulary web page). Batch metadata is used for separating different datasets and facilitates downloading and uploading data in batches with additional batch information and supplementary files.

Additional flexibility, built into the database infrastructure using an RDF framework, will enable the system to host different taxonomic groups, which might require new metadata features (Jebbia et al. 2023). By the combination of these features, along with FAIR (Findable, Accessable, Interoperable, Reusable) principles, and reproducibility, we provide Artificial Intelligence Readiness (AIR; Long and Romanoff 2023) to the dataset.

Fish-AIR provides an easy-to-access, filtered, annotated and cleaned biological dataset for researchers from different backgrounds and facilitates the integration of biological knowledge based on digitized preserved specimens into ML pipelines. Because of the flexible database infrastructure and addition of new datasets, researchers will also be able to access additional types of data—such as landmarks, specimen outlines, annotated parts, and quality scores—in the near future. Already, the dataset is the largest and most detailed AI-ready fish image dataset with integrated Image Quality Management System (Jebbia et al. 2023, Wang et al. 2021).

Keywords

biodiversity informatics, data management, machine learning, artificial intelligence, data wrangling

Presenting author

Yasin Bakış

Presented at

TDWG 2023

Funding program

- NSF Harnessing the Data Revolution Institute Grant #2118240 (Imageomics)

- NSF Office of Advanced Cyberinfrastructure Grant #1940322 (BGNN)

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Bakış Y, Wang X, Jr HB (2021) Evaluating the image quality of digitized biodiversity collections' specimens. Great Lakes Bioinformatics Conference, May 10 2021.

- Bart H, Greenberg J, Karpatne A, Mabee P, Maga A (2021) Biology-guided neural networks (BGNN) for discovering phenotypic traits. Society for Integrative and Comparative Biology 2021 Virtual Annual Meeting, January 3 2021.
- Jebbia D, Wang X, Bakis Y, Bart H, Greenberg J (2023) Toward a Flexible Metadata Pipeline for Fish Specimen Images. Metadata and Semantic Research175-190. https://doi.org/10.1007/978-3-031-39141-5_15
- Karnani K, Pepper J, Bakış Y, Wang X, Bart H, Breen D, Greenberg J (2022) Computational metadata generation methods for biological specimen image collections. International Journal on Digital Libraries https://doi.org/10.1007/s00799-022-00342-1
- Leipzig J, Bakış Y, Wang X, Elhamod M, Diamond K, Dahdul W, Karpatne A, Maga M, Mabee P, Bart H, Greenberg J (2021) Biodiversity Image Quality Metadata Augments Convolutional Neural Network Classification of Fish Species. Metadata and Semantic Research3-12. https://doi.org/10.1007/978-3-030-71903-6_1
- Long S, Romanoff T (2023) AI-Ready Open Data. https://bipartisanpolicy.org/explainer/ai-ready-open-data/. Accessed on: 2023-8-22.
- Pepper J, Greenberg J, Bakış Y, Wang X, Bart H, Breen D (2021) Automatic Metadata Generation for Fish Specimen Image Collections. bioRxiv https://doi.org/10.1101/2021.10.04.463070
- Wang X, Bakış Y, Jr. HB (2021) Gathering specified and standardized image quality metadata through a cyber infrastructure. Great Lakes Bioinformatics Conference, May 10 2021.