SC/68B/PH/01

A Cross-reference of Flukebook and Happywhale Platforms

Paula A. Olson (compiler), Drew Blount, Ted Cheeseman, Jason Holmberg, Gianna Minton

INTERNATIONAL WHALING COMMISSION

A Cross-reference of Flukebook and Happywhale Platforms

Paula A. Olson (compiler)[1], Drew Blount[2], Ted Cheeseman[3], Jason Holmberg[2], Gianna Minton[4]

[1]*Marine Mammal and Turtle Division, Southwest Fisheries Science Center NMFS/NOAA, 8901 La Jolla Shores Drive, La Jolla, CA 92037 USA*

[2] *WildMe*

[3] *Happywhale*

[4] *Megaptera Marine Conservation/Arabian Sea Whale Network*

Paula.Olson@noaa.gov

Background

In 2019 at SC68A, there was discussion in SH sub-committee by members who wished to see a side-by-side comparison of the two automated photo-ID recognition software platforms, Flukebook and Happywhale. Both use algorithms to match individual humpback whales from photographic catalogues.

Flukebook was developed under the Wildbook platform, which was initially focused on matching and archiving individual whale shark (*Rhincodon typus*) and giant manta ray (*Mobula birostris*) photos and then later cross-applied to cetacean data. Since initial development of computer matching algorithms for humpback whales, Flukebook has grown with algorithms developed or integrated for several additional cetacean species. See also Blount *et al.*, 2019 (IWC SC/68A/SH07).

Happywhale uses automatic image recognition software for humpback whale flukes that has achieved a high rate of accuracy and success in matching. Originally Happywhale was focused on humpback whales in the South Pacific but it has expanded its efforts to populations world-wide. See also Cheeseman and Southerland, 2018 (IWC SC/67B/PH05).

Here we present a comparison table of the two platforms. It is considered a draft table and we look to members of the *Ad hoc* Working Group on Photo-identification to refine the table so that it provides a useful reference for the Scientific Committee.

Table cross-referencing features of Flukebook and Happywhale. Organized in three general categories: I) Access, Users, Uploading; II) Matching, Data, Outputs; III) Overview.

	Flukebook	**Happywhale**
I. Access, Users, Uploading		
Link:	**www.flukebook.org**	**www.happywhale.com**
Access: Open to all? Need account?	Open to all for data submission. Data curation requires administered login account.	Open to all, accounts self-created for submitting users
Can users see data from other contributors?	Only with express permission granted at the Organization or pair-wise User level after login.	Yes, by default. Users can set use permissions and privacy at an image level
Data protection measures?	Yes: multiple. Secure login to view data. HTTPS data transmission. Firewall. GDPR compliance. Data only visible to permitted Users or Organizations of Users after login. Data can be shared with other users as read-only or as editable. All data rights remain with contributor.	Yes: (1) with user-set creative commons licensing permissions, and (2) with image level firewall privacy based on permissions groups for defined collaborations, keeping any data desired visible only within a defined set of users. HTTPS data transmission. GDPR and California CPA compliant.

	Flukebook	Happywhale
I. Access, Users, Uploading cont.		
If yes for data protection, how does it work?	Data visibility is currently limited to researchers with login access. After login, other users' data is visible only with express permission and can be User-managed. Organization membership can also control read-only and editable access to data. Data search and export (multiple formats) rely on record-level security checks to ensure only permitted data can be reviewed and exported. Where data crosses an unpermitted boundary, Users are prompted to initiate collaboration or are blocked.	Creative commons settings: https://happywhale.com/account (bottom of page; requires being signed into a self-created account first) Privacy groups: Defined privacy groups allow data access within collaborator groups created at an admin level. User can set privacy of any data upon upload.
How do users upload historical data sets?	In Wildbook standard Excel format: https://www.flukebook.org/import/instructions.jsp Custom data blending and import available quickly to support complex legacy data sets.	https://happywhale.com/submitMedia or by batch upload with metadata and/or spreadsheet based attribute data, with integrated rapid image recognition, quality scoring and image attribute tagging processes
How do users upload individual whales?	https://www.flukebook.org/submit.jsp	Typically via https://happywhale.com/submitMedia Advanced users (engaged citizen scientists and research users) can access an advanced submission process to open a richer upload process.
Can users modify their own data once it is uploaded?	Yes for all users.	Personal account settings, yes, all users. Citizen science users: No data editing apart from encounter level commenting (also, effectively creating social media functionality and crowd-sourced data curation). Research users can access and administer their data, yes.
Can users access their own data once it is uploaded?	Yes. Full access, including Encounter Search, Individual Search, Occurrence Search, record-level editing, and security-filter data export. Search display includes sortable/filterable table, map display, calendar display, analysis display, and data export display.	Yes, via dedicated user pages, dedicated organization pages and map and attribute-based search tools. Some advanced search tools and data fields are accessible only at the research collaborator and admin level
Are there costs associated with uploading or maintaining data on the platform?	No	No

	Flukebook	Happywhale
II. Matching, Data, Outputs		
For which species (and which features of those species - flukes/dorsal fins/head callosity patterns) are automated/AI algorithms available for matching?	Matching/identification is part of a fully automated pipeline that includes feature detection, viewpoint prediction, and background subtraction (Parham *et al*., 2018). • *Megaptera novaeangliae* - Fluke (x3 algorithms; feature-based[1] and deep learning) • *Physeter macrocephalus* - Fluke (x2 algorithms; feature based) • *Tursiops truncatus* – Fin (x2 algorithms; feature-based and deep learning) • *Tursiops aduncus* – Fin (x2 algorithms; feature-based and deep learning) • *Delphinus delphis* – Fin (x2 algorithms; feature-based and deep learning) • *Eubalaena glacialis* – Callosity pattern (aerial) (x1 algorithm; deep learning) • *Eubalaena australis* – Callosity pattern (aerial) (x2 algorithms; feature-based and deep learning) • *Stenella frontalis* – Flank patterning (x1 algorithm; feature-based)	Humpback whale flukes
What is accuracy rate of matching for each algorithm used on the platform?[2] -	Varies by algorithm, species, and especially data quality. Examples: • Humpback Flukes – Kaggle 7 – Top 1: 93%, Top-5: 97% against 4511 whales (6700+ images) from the Cascadia Research Collective Pacific catalog. • Humpback Flukes – CurvRank – Top 1: 80% (Weideman *et al*., 2017) • Bottlenose dolphin fins – Top 1: 95%, 400+ individuals (10,000+ photos) from the Sarasota Dolphin Research Project catalog. • North Atlantic Right Whales [3] (Bogucki *et al.*, 2019) – Deepsense – Top 1: 87%, Top-5: 95%	97% - 99% of potential matches found with good to high quality images in first proposed result, > 90% of potential matches found in top 5 proposed results with poor to moderate quality flukes, based on an analysis of match results with 4753 fluke ID images from known whales from Glacier Bay National Park.

[1] Feature-based systems allow for immediate matching of individuals and adding new individuals using extracted features without requiring external deep learning model retraining.

[2] Ideally there would be a standard measure that could be used for each platform and each algorithm for each species/feature - e.g. on 96% of matching runs, the top 10 photos returned from a search include the match provided by manually matched training data.

	Flukebook	Happywhale
II. Matching, Data, Outputs cont.		
Are photos automatically matched to all photos on the platform? Or are users in control of which catalogues/datasets their contributions/uploads are matched to?	Users can match versus all photos of the same species (or same genus where exact species is not known) or against any regional subset or combination of regions.	All photos, currently 35,000 individuals. Subsets can be configured
Can users engage in 'catalogue shares' or matching with other users?	Yes	Yes
If yes to catalogue shares, how does it work?	Users can multi-select from the global catalog, matching against individuals from those study sites using feature-based matchers (e.g., finFindR, CurvRank). For pre-trained ML models (e.g., Deepsense), matching is against the global catalog (or quality subset the model was trained on). All users can match against all individuals, but the display of potential matches is partially limited to protect data ownership and to prompt for User-initiated collaborations and broader data sharing inside Flukebook. Example video: https://www.youtube.com/watch?v=caZ1rDacrXo	Subsets and custom sets can be configured for specific duplicate searches within collaborator catalogs. However it is usually more effective, more efficient and more instructive to match against all known whales.
Can the Platform also be used to store and manually match photos of species for which AI matching algorithms are not yet available?	Yes, and Flukebook is actively being used this way for Orcas and fin whales. Because Flukebook has an integrated AI pipeline for multi-species detection and identification (supported by a dedicated, full-time machine learning team), manual data storage and ID curation in Flukebook is in itself preparation for creating, training, testing, and independent validation of future AI algorithms as Flukebook pushes this data to its backend machine learning platform (Wildbook). We can also cross-apply matching techniques from one species to another (e.g., *Tursiops truncatus* to *Tursiops aduncus*) via simple configuration changes, creating incredibly fast onboarding of new research teams, catalogs, and species.	Yes. Happywhale has become a fruitful and accessible source for opportunistic sightings, currently with sighting records of over 80 species. The platform serves as a single source for curating the chaos of citizen science data made available in an effective standardized format for research use. The user experience has been sufficiently positive that participation is actively promoted on many expedition cruise and whale watch vessels, to the extent of web connectivity being a free unthrottled site on some vessels where internet access is otherwise paid. Dedicated volunteer ID work has resulted in (currently) 238 ID'd individual killer whales of eight ecotypes.
Quality Control: Does the Platform have a way to score photos on quality and/or distinctiveness in order to meet the photo-ID recommendations of Friday *et al*. 2008 and Urian *et al*. 2015?	Yes	Yes

	Flukebook	Happywhale
II. Matching, Data, Outputs cont.		
Does the Platform provide an export function for the users' data to MARK?	Yes	Yes
Export function for R?	Yes - https://cran.r-project.org/web/packages/RWildbook/index.html	Yes
Export function for GIS for mapping?	Yes, KML and Shapefiles as well as in-browser visualization.	Yes. The platform is a simple web-based GIS as well
Export function for Excel?	Yes	Yes
Export function for SOCPROG?	Yes	On request
Export functions for any other applications?	Yes (e.g., Excel, CSV, PDF picture book), and customizable	Automated ongoing export to OBIS-SEAMAP for accessible shareable data, Fluke catalog PDF sorted by pigment category, CSV (Darwin core format)
Can the Platform be used to store sighting data for cetacean sightings that do not have associated photos?	Yes, especially for genetics (see GeneGIS project)	Yes. We are curating sighting data from various sources including our marine mammal sighting mobile app (currently primarily in use by expedition guides on polar cruise tourism vessels)
Can the Platform be used to filter/analyse data on cetacean sightings that do not have associated photos?	Yes	Yes
If yes for no-photo cetacean sightings, please list data fields that are included (e.g. date, time, location, sea state, associated effort, group size, group composition, human activity in proximity to sighting etc.).	Date, time, GPS, study site, verbatim location description, genotype, haplotype, submitters, visual code, species, measurements, tissue samples and related metadata, data owner, satellite tag metadata, Survey track, Survey vessel, individual ID (e.g., if genetically determined), genetic sex, sex, familial relationship (e.g., mother-calf), social relationship (e.g., pod), country, scarring, behavior, life stage, work flow state, record metadata (e.g., date created), and more.	All non-media specific attributes can be stored for records at the sighting and individual level without media. Data such as sea state, effort, GPX vessel track, platform, participating observers etc are recorded at the survey level.
Can the Platform accommodate other types of data for individuals?	Yes, configurable.	Yes, configurable
If yes, what are other types of data? Please list:	Robust support for genetic, toxicology and social data. Genetic: tissue\biological sample type, haplotype and genotype. Genetic sex determination. Tagging: stores metadata for related physical tags (e.g., satellite) Toxicology: stores multiple chemical measurements on tissue\biological samples.	For migratory species we record multiple IDs across multiple reconciled catalogs. Individual level attributes (such as fluke pigment category, haplotype, genotype etc, configurable) are stored at the individual level whereas event-based data (ie sampling, tagging) are stored at an encounter level.

	Flukebook	Happywhale
III. Overview		
What are the features of the Platform most likely to appeal to cetacean researchers?	*Data Security and Privacy:* While supporting citizen science contribution of data, Flukebook.org considers the data security and privacy of researchers as its primary User scenario. Pair-wise User collaboration and Organization-based collaboration support real-world research networks (e.g., Indocet, ASWN) and allow for concrete and lasting reconciliation across multiple catalogs for multiple species. *Multi-lingual support for a global community.* *Professional staffing and support:* Flukebook Users have access to dedicated, experienced machine learning and software experts on the Wild Me team (https://www.wildme.org/contact), and we have recently launched our professionally-staffed support community at https://community.wildbook.org. Flukebook is open source and staffed solely by non-profit technologists, allowing us to take a trusted, non-competing stance when dealing with the data security of researcher across the globe. *Unparalleled, Automated Computer Vision Pipeline for Multiple Species:* Flukebook.org uses the broader Wildbook.org infrastructure that supports algorithm- and machine learning-based solutions for detection and identification of multiple species across multiple platforms (e.g., whaleshark.org, mantamatcher.org, giraffespotter.org). This open source pipeline is continuously improved and ensures that new techniques can be quickly integrated as "plugins" and become available to researchers across the globe through a simple browser.	The feedback we receive is appreciation of the rapid high-quality photo ID automation for humpback fluke matching, ease of enabling collaboration (with various mixtures of data privacy) and for humpbacks as well as other species, provisioning of otherwise inaccessible data in a functional standardized format. Platform ease of use, professional quality interface and accessibility also has been complemented. Especially relevant in the case of a migratory species studied regionally, the platform is reconciling catalogs and image libraries to the extent of currently accommodating over 90 naming protocols; we have several whales with eight different catalog IDs. We are well integrated within the marine mammal science community, actively participating in archiving and contributing relevant data to conservation science in a not-for-profit context. For the research community, this means a whale scientist is involved and engaged with data management, adding critical context and quality control oversight. The platform creates a unique channel to connect with citizen scientists, as a source of data and/or communication with engaged marine mammal enthusiasts.

	Flukebook	Happywhale
III. Overview cont.		
What are the features of the Platform most likely to appeal to citizen scientists?	At Wild Me, we have 18 years of experience managing publicly contributed data, starting with whale sharks and manta rays and now moving on to multiple terrestrial and marine species. Citizen science data is only of scientific benefit if it can be effectively utilized by the researchers who can use it. To that end, we start with a researcher-first use case and then add citizen science to augment structured studies. For citizen scientists, we support: • Direct contribution of data: https://www.wildme.org/contact/ • Automated emails confirming: data submission, ID assignment, future resights. • Automated ID pipeline, allowing researchers to use and confirm matches faster, generating automated notifications to citizen scientists faster.	Globally it is unprecedented for citizen scientists to be able to submit a photo of an individual whale and rapidly receive notification of the whale's ID, as well as further notifications of resightings of 'their' whales. We have seen substantial organic growth in platform use based on positive user experiences, creating data streams in otherwise data scarce environments and for the citizen scientist, a strong feeling of being able to meaningfully contribute to conservation science. "Telling passengers who [the whales] are, giving them a background story and talking about them as an individual definitely engages the public, and I believe it creates a sense of stewardship and deeper appreciation for our ocean environments." ~ Jessica Roame, Marine Education Specialist, Newport Landing ""Wow, its hard for me to express by email just how satisfying and fun it is to know that "my" whales are new to science. I'm not sure what it is about it, but seeing my flukes located on the map is really fulfilling and gives me a connection to them I've never experienced before. I'm also psyched to see that passengers have also been contributing. This "Happywhale thing" is working!" ~ Eric Carr, polar expedition guide
What are future developments/improvements intended for the Platform?	• "Graph algorithm", using ML for internal error checking and dynamic reassignment of IDs to promote more accurate catalogs. • Adding additional species and ID techniques in AI (e.g., "triplet loss networks") • User interface v2 underway • Support for languages beyond the existing English, Spanish, and French. • Mobile app integration through upcoming Ocean Alert	- Expanding mobile app implementation for marine mammal sightings, from opportunistic to full featured MMO sighting survey - Multi-species automated ID - Multi-lingual support - Ongoing development, improvement and adaptation

	Flukebook	Happywhale
III. Overview cont.		
Overall: Platform strengths	• Robust, modern platform (Java/Python) proven across many species and implementations with daily improvement and update • Free to use • Multi-species, fully automated detection and ID pipeline, including best-in-class algorithms for humpbacks, dolphins, and right whales • Professional support and staffing by software and AI experts (8) • Non-profit and open source • Multi-lingual for global audience • Comprehensive data search and novel data visualization. • Active support from and use by multiple organizations across	- High quality cohesive and modular fully refactored up-to-date coding - Free - Accessible. For example, with one click a user can see everywhere 'their' whales (or a selection whales based on any search criteria) have been recorded. - Best-in-class automated ID for humpback flukes - Effective data management of cit sci and research collaboration data, integrated - Responsive and collaboration-friendly focus on marine mammals while benefitting from research software and citizen science platform development in collaboration with the non-profit Polar Citizen Science Collective
Overall: Platform weaknesses	• As a relatively large effort, we require significant annual funding from our patrons (e.g., Microsoft, Gordon and Betty Moore Foundation) • We are redesigning the user interface to simplify the self-service data submission and ID workflows, especially for multi-animal, multi-photo use cases for more social species. • We are experiencing rapid growth and need to create a more sustainable set of platform training materials to support rapid onboarding of new users (i.e. better documentation) • We are currently reworking our AI architecture to improve self-service, multi-user response times on automated ID matching, which is experiencing a high volume of usage.	We are a small team, and we are not self-promotionalists. Based on for example the IWC 2019 report that stated the concern that all data is open access, capacity and capabilities are poorly understood. The site needs text and interpretive content updates.

References

Blount, D., Homberg, J., Minton, G., Kahn, C. 2019. Flukebook - Recent advances for cetacean photo identification and data archiving including automated fluke matching. Paper IWC SC68A/SH07 presented to the IWC Scientific Committee.

Bogucki, R., Cygan, M., Khan, C.B., Klimek, M., Milczek, J.K., and Mucha, M. 2019. Applying deep learning to right whale photo identification. *Conservation Biology* 33, 676–684.

Cheeseman, T. and Southerland, K. 2018. Happywhale progress report 2017-2018. Paper IWC SC/67B/PH05 presented to the IWC Scientific Committee.

Friday, N.A., Smith, T.D., Stevick, P.T., Allen, J., and Fernald, T. 2008. Balancing bias and precision in capture‐recapture estimates of abundance. *Marine Mammal Science 24*(2), 253-275.

Parham, J., Stewart, C., Crall, J.P., Rubenstein, D., Holmberg, J., and Berger-Wolf, T. 2018. An Animal Detection Pipeline for Identification. 1075-1083. 10.1109/WACV.2018.00123.

Urian, K., Gorgone, A., Read, A., Balmer, B., Wells, R.S., Berggren, P., Durban, J., Eguchi, T., Rayment, W., and Hammond, P.S. 2015. Recommendations for photo‐identification methods used in capture‐recapture models with cetaceans. *Marine Mammal Science 31*(1), 298-321.

Weideman, H.J, Jablons, Z.M., Holmberg, J., Flynn, K., Calambokidis, J., Tyson, R.B., Allen, J.B., Wells, R.S., Hupman, K., Urian, K., and Stewart, C.V. 2017. Integral Curvature Representation and Matching Algorithms for Identification of Dolphins and Whales. doi: 10.1002/ar.23650