Meeting Report for the Phenoscape TraitFest 2023 with Comments on Organising Interdisciplinary Meetings

Jennifer C. Girón Duque[‡], Meghan A Balk[§,|], Wasila Dahdul[¶], Hilmar Lapp[#], István Mikó[¤], Elie Alhajjar[«], Brenen Wynd[»], Sergei Tarasov[^], Christopher Lawrence[˅], Basanta Khakurel[»], Arthur Porto[¦], Lin Yan[˾], Isadora E Fluck[ˢ], Diego S Porto[^], Joseph N Keating[¢], Israel T Borokini[˾], Katja Chantre Seltmann[ℓ], Giulio Montanaro[^], Paula Mabee[ঌ]

‡ Natural Science Research Laboratory, Lubbock, United States of America
§ Naturhistorisk Museum, Universitetet i Oslo, Oslo, Norway
| National Ecological Observatory Network, Battelle Memorial Institute, Boulder, United States of America
¶ University of California, Irvine, Irvine, United States of America
# Duke University, Durham, NC, United States of America
¤ University of New Hampshire, Durham, United States of America
« RAND Corporation, Arlington, United States of America
» Southeastern Louisiana University, Hammond, United States of America
^ Finnish Museum of Natural History, Helsinki, Finland
˅ Princeton University, Princeton, United States of America
¦ Louisiana State University, Baton Rouge, United States of America
˾ University of California, Berkeley, Berkeley, United States of America
ˢ University of Florida, Gainesville, United States of America
¢ University of Bristol, Bristol, United Kingdom
ℓ University of California, Santa Barbara, Santa Barbara, United States of America
ঌ National Ecological Observatory Network, Battelle, Boulder, United States of America

Corresponding author: Jennifer C. Girón Duque (entiminae@gmail.com)

Academic editor: Gail Kampmeier

Abstract

The Phenoscape project has developed ontology-based tools and a knowledge base that enables the integration and discovery of phenotypes across species from the scientific literature. The Phenoscape TraitFest 2023 event aimed to promote innovative applications that adopt the capabilities supported by the data in the Phenoscape Knowledgebase and its corresponding semantics-enabled tools, algorithms and infrastructure. The event brought together 26 participants, including domain experts in biodiversity informatics, taxonomy and phylogenetics and software developers from various life-sciences programming toolkits and phylogenetic software projects, for an intense four-day collaborative software coding event. The event was designed as a hands-on workshop, based on the Open Space Technology methodology, in which participants self-organise into subgroups to collaboratively plan and work on their shared research interests. We describe how the workshop was organised, the projects developed and outcomes resulting from the workshop, as well as the challenges in

bringing together a diverse group of participants to engage productively in a collaborative environment.

Keywords

biodiversity, phenotype, biodiversity informatics, knowledge base, ontologies, Phenoscape

Introduction

Trait data that are amenable to computational data science, including computation-driven discovery, remain relatively new to science. Efficiently repurposing, integrating and mining the vast stores of trait data have long been hampered by the limited amount of data accessible in standard formats and by the challenges involved with enabling machines to compute data that are largely recorded in natural language. A variety of resources have been developed to address these challenges, including powerful knowledge representation technologies (Mungall et al. 2007), shared domain ontologies (Haendel et al. 2014), very fast machine reasoners (Kazakov et al. 2013, Glimm et al. 2014) and various tools accelerating the construction of large databases of ontology-linked phenotype data (knowledge bases; Thessen et al. (2020)). The results emerging from these advances provide new opportunities for computation-driven data science with trait data. For biodiversity research, a unique resource of data science-enabled trait data is the Phenoscape Knowledgebase (KB), a datastore of vertebrate morphological traits linked to terms drawn from formal ontologies and, thus, represented in a structured and fully computable form.

Since 2007, the NSF-funded Phenoscape project*[1] has focused on transforming natural language descriptions of comparative traits into computable formats using knowledge representation and discovery technologies, in particular, ontologies and semantic technologies (Dahdul et al. 2010, Dececchi et al. 2015, Mabee et al. 2018), with the goal of rendering the descriptions interoperable and recombinable (Mabee et al. 2012) and linking phenotypes to candidate genes (Edmunds et al. 2015) and gene networks(Fernando et al. 2023). The project's outcomes include the creation of resources, such as the Phenoscape KB (Balhoff and Phenoscape Project Team 2016), tools for curating and staging data and ontologies (Balhoff et al. 2010, Balhoff et al. 2014), as well as visual and programmatic query interfaces.

The Phenoscape KB is an online resource that contains evolutionarily-relevant phenotypic trait data from over 250 comparative morphology studies to date and with a primary focus on the vertebrate fin-to-limb transition and comparative fish morphology (Mabee et al. 2018). The KB includes natural language phenotype descriptions annotated with terms from formal ontologies. By using ontologies for morphological, spatial and other knowledge domains, the KB links genotypes to phenotypes from genetic perturbation studies of model organisms, such as zebrafish, the African clawed frog, and

mouse, as well as to human genetic disease. These semantic linkages enable the discovery of relationships between traits that were previously unknown using traditional methods, as the logic underlying ontologies can infer relationships (Edmunds et al. 2015, Manda et al. 2015, Mabee et al. 2018).

The KB also offers an application programming interface (API) that enables exploration of connections amongst traits and between taxonomic groups. These include access to machine-reasoning-based algorithms, such as presence/absence reasoning for characters and states that are implied by, but not necessarily asserted in, original studies (Dececchi et al. 2015). For example, a phenotype described in literature for the count of supporting rays in the dorsal fin implies the presence of a dorsal fin in the organism and this latter phenotype is inferred by machine-reasoning in the Phenoscape KB. The KB also provides access to algorithms that find candidate genes for evolutionary phenotypes by connecting ontology-annotated evolutionary phenotypes for vertebrates with model organism genetic data; for example, a KB query for the phenotype 'scale, absent' will return candidate genes eda and edar from semantically similar gene phenotypes in zebrafish (Mabee et al. 2018, Edmunds et al. 2015).

Phenoscape's current subproject, Semantics for Comparative Analysis of Trait Evolution (SCATE), is developing tools that use the KB's data and computational capabilities to assist in analyses of trait evolution (Tarasov et al. 2019, Porto et al. 2022). This is achieved by providing comparative trait analysis tools with easy access to KB data, notably in the form of dependency matrices representing prior knowledge about anatomy (e.g. the presence of a dorsal fin ray is dependent on the presence of a dorsal fin) and semantic similarity matrices, based on semantic distance between phenotypes (Lapp et al. 2022, Porto et al. 2023).

At present, the Phenoscape KB primarily focuses on semantically-encoded vertebrate phenotypes, while the SCATE project centres on developing ontology-enhanced tools and techniques for assisting in phylogenetic comparative analyses. The infrastructure and toolset developed by Phenoscape and SCATE hold enormous potential for broader applications beyond vertebrate systems. To fully leverage this potential, broader adoption of Phenoscape/SCATE resources is needed by users and taxonomic experts from diverse communities. However, writing semantic descriptions of traits is a bottleneck for ontology-based knowledge bases (Dahdul et al. 2015). To expedite creation of semantic descriptions, machine-readable text-mining applications are needed to extract key words and these tools need to be better aligned with automated image recognition tools for extracting phenotypic information and metadata. Doing so will enable more comprehensive and efficient analysis of phenotype data across a wider range of organisms and scientific domains.

In response to these challenges, Phenoscape/SCATE hosted TraitFest 2023, a global workshop held at the Renaissance Computing Institute (RENCI) in Chapel Hill, North Carolina, from 23-26 January 2023. The primary objective of the event was to engage potential users and contributors to the data and infrastructure provided by Phenoscape/ SCATE, as well as developers of methods, especially in comparative phylogenetics and

related fields. We aimed to include users whose research could benefit from computable semantics-based capabilities and whose taxonomic communities have already developed the necessary baseline knowledge representation infrastructure not currently present in the Phenoscape KB, such as for Arthropoda (Yoder et al. 2010, Balhoff et al. 2013, Girón et al. 2023). We also sought to engage those interested in developing tools or workflows, particularly those using machine-learning for natural-language and image processing. The goal was to foster a diverse community of users and contributors who can leverage the full potential of the Phenoscape/SCATE infrastructure and data for their research.

Here, we describe our approach to organising the event as a collaborative hands-on unconference-style workshop, based on the Open Space Technology (OST) methodology (Owen 2008). In the OST method, participants self-organise into groups to collaboratively tackle self-selected projects. We also describe challenges and successes in bringing together participants with completely different skills, areas of expertise and interests, to work together in a short amount of time to generate project ideas and contribute meaningfully and productively during the workshop. We focused on devoting time during the workshop to learning activities (e.g. bootcamps) to share specialised knowledge, such as ontologies or machine-learning. We also describe the outcomes of the projects developed during the workshop.

Organisation

To facilitate inclusive decision-making and task sharing, an organising committee was assembled to include individuals broadly resembling the anticipated audience for the event. The full list of participants, including the organisers with their fields of expertise and interests can be found on the workshop wiki*[2]. Likewise, for the event, we aimed to gather a diverse group of individuals, including diversity in expertise from relevant fields like evolutionary biology, ecology, biodiversity science, biomedical sciences, bioinformatics, data science, machine-learning and computer science. We sought researchers, software/tool developers, training/documentation specialists and visual and data interaction designers.

We invited people two ways: targeted invitations sent by members of the organising committee and by an open call for applications. We posted an open call for participation*[5] on Twitter and biology-related mailing lists (i.e. evoldir and ECOLOG-L). Our criteria for selecting participants from the pool of applications included the following: 1) expertise in Insecta, ontologies or machine-learning; 2) motivation to learn or use Phenoscape in research; 3) contribution to diversity such as demography, perspective or background; 4) contribution to knowledge in invertebrates or computational expertise. The 18 invited participants (including one remote) and six members of the organising committee and two Phenoscape participants included international participants from Finland, the United Kingdom and Italy.

In advance of the workshop, participants were asked to complete several tasks that would help them prepare for the meeting and maximise the time available at the meeting for organising groups and working on group projects. It was critical to familiarise participants with the very wide range of interests and expertise held by the larger group and to provide a platform for participants to propose ideas ahead of the workshop. We set up a GitHub repository,*[3] which included a wiki*[4] for participant materials and information. An email, sent two weeks before the start of the workshop, outlined specific tasks for participants to complete before the start of the meeting. These included filling in their information, including expertise and interests, in the participants' table on the wiki page. We also asked that participants create and/or contribute to an issue on the repository discussing a research idea, funding opportunity, data source(s), bootcamp and/or technology to be developed. The issue tracker helped the organisers understand the participants' interests, skills needed for different ideas, knowledge gaps and where to have bootcamps. On the wiki, we also had a code of conduct,*[6] various resources and an agenda.*[7] Participants were advised that they should actively contribute or learn and to conceive project ideas that would take advantage of the workshop's unique opportunities to collaborate with fellow participants (including organisers).

We also set up a Slack channel for general information (SCATE TraitFest) and invited all participants and relevant RENCI meeting logistics personnel. Participants were encouraged to use Slack before the meeting, for example, to arrange shared transportation from the airport, during the meeting to share resources and after the meeting to continue collaborations.

Overview of activities

The meeting was organised around the Open Space Technology (OST) concept in open science (Owen 2008). In the OST methodology, participants self-organise into groups, based on common interests and work together towards shared goals or outcomes. Our meeting had such a broad research audience that predefining tasks or topics for the workshop was nearly impossible, which is why we opted for OST. With OST, participants have the freedom to plan their own activities, in contrast to pre-planned conferences where organisers schedule speakers in advance. This approach empowers participants to take control of the event and reduces the need for extensive pre-planning. In our implementation of OST for this event, the organisers transitioned into the role of participants as the event neared.

The workshop agenda*[7] was developed to prioritise work time and learning opportunities. After introductions and a keynote presentation by Tanya Berger-Wolf (Ohio State University) about the Imageomics Institute, participants engaged in two rounds of Open Space group pitches on the first day. In the first round, participants pitched an idea for a research project, funding proposal or technological development. These were written on large post-it notes, hung around the room and subsequently grouped by topic. The next round required participants to identify which ideas they were interested in. Participants needed to choose one main idea and an idea needed more than two people

to move forward, with an eye towards a maximum of eight ideas (based on the number of people and time available) to develop during the workshop. This required a round of conversations between participants who did not pitch an idea and those who did, as well as between leaders of ideas to potentially consolidate and synergise plans.

The final, self-assembled groups*[11] then developed projects with potential products or available datasets in mind (see Results). Participants were guided in their work by following predefined rules that were sent in an email: self-organise, share knowledge and be proactive and productive. We worked on the projects and reported on progress at the end of each day. Groups were encouraged to create shared resources that were FAIR (Findable, Accessible, Interoperable, Reusable; Wilkinson et al. (2016)) and open, such as GitHub repositories. Groups created channels on Slack to aid in sharing information and organising. At the end of the workshop, we provided a final report that is encapsulated in the results section of this document.

Ad hoc bootcamp sessions for training or information sharing during the first couple of days of the workshop were also encouraged. Bootcamps were participant-requested, short, informal sessions led by participants in their area of expertise. Bootcamp topics included: ontologies and Phenoscape KB tools led by J. Balhoff; machine-learning from images led by A. Porto; curation of matrix-based phenotypic descriptions using Phenex (Balhoff et al. 2014)*[8] led by W. Dahdul; Phenoscript*[9] for the creation of semantic species descriptions led by S. Tarasov; phylogenetic comparative methods led by J. Keating; and the development of the Ontology for the Insect SkeletoMuscular System*[10] led by J. Girón.

A few days after the workshop, we shared a survey to learn the perceptions of the participants and how successful and productive the methodologies were.

Results

The eight projects pursued by participants during the workshop are summarised below.

Project I: Images to Traits

Problem to solve: Images are constantly being generated in biodiversity research as an important source of information about characteristics (traits) of organisms and in the ongoing digitisation of biological collections. However, extracting trait information from those images is generally time-consuming, if even possible, such that humans cannot keep up with the volume of images generated daily. Thus, the trait information encoded in those images remains 'dark' (Marshall et al. 2018) and is never digitised to machine-processable data. This group was interested in exchanging information about using images as a source of trait information and learning the state-of-the-art in image processing for machine-learning tasks (such as trait extraction), data and processing documentation and metadata standards.

Approach: The group of people included a botanist (I. Borokini), an entomologist (J. Girón) and two informaticians (B. Altintas and X. Wang). B. Altintas and X. Wang have generated pipelines for trait annotations in fishes, along with public views of these data. This group determined that the first thing needed was a paper compiling resources used for the different aspects of generating and using images, especially for machine-learning approaches. They spent time learning from each other the "why" and "how" they generate and process images for their work.

Results of the workshop: Along with the participation of several other workshop participants, they put together a document with a list of resources, tools and considerations when generating and using images for research in biodiversity. This document is available as a Google Doc and linked in the workshop's GitHub repository*[11] . They also started drafting the manuscript to tie together all these resources, but did not manage to pass beyond the introduction, as the nature of the workshop itself diverted the attention of group members to other projects, where their expertise was needed.

Future directions/plans/recommendations: Their plan is to keep working on the resources paper, from the point of view of both the biologist, who generates and uses images for particular purposes (often illustrative) and from the informatician's point of view, who knows the standards and processes needed to extract data out of those images in a format amenable for downstream computation. In addition, a discussion with a broader group of participants around the topic of documentation and metadata standards for image annotations became an after-workshop endeavour (M.A. Balk, W. Dahdul, J. Girón, A. Porto). This sub-group wants to engage the broader biodiversity standards and bio-ontologies communities in this conversation about documentation, metadata and standards (including AudioVisual Core *[39] and tracking changes to an image in preparation for inclusion in AI pipelines) and possibly publish, at least an opinion piece, about the discussions had on the topic during the workshop.

Project II: Automated quantification of video and image files

Problem to solve: Research in the biological sciences is driven by a desire to provide important context and mechanisms for phenomena seen across the natural world. A common approach across and between disciplines is to assess the presence/absence of a trait (morphological, behavioural), by measuring and comparing the trait(s) of interest to establish a correlation. Regardless of the organism in question, the ability to generate reliable and readily comparable (homologous) measurements is critical to better understand the processes that influence the evolution of our ecosystems. However, data collection is often the most time and resource intensive period of many biological studies, due to a lack of access to specimens (e.g. lack of travel funds) and/or time required to learn measurement protocols and to determine viable, useful measurements within and across taxa. Data can be obtained from recorded images or videos from the organisms of interest. The goals of this project are to expedite the process of data acquisition and implement machine-learning approaches to automate the collection of measurements from large datasets. With faster, but still reliable data collection, researchers can focus more on processing and modelling.

Approach: This group was split into behavioural (video data) and morphological (image data) subgroups. As the primary focus of this group was acquisition of data, the group included many of the most junior individuals at the workshop, those who are still acquiring data for various research projects, including four graduate students (B. Khakurel, C. Charpentier, C. Lawrence, L. Yan) and one postdoctoral research fellow (B. Wynd). For each subgroup, the primary approach and goal of the workshop was to develop pipelines using existing tools to quantify morphology and behaviour. As these workshops tend to be short and communication can dwindle afterwards, the group prioritised the development of the pipelines to facilitate research and allow for continued development after the conclusion of the workshop. An additional, but secondary goal, for the image-subgroup was to evaluate the utility of point- (landmark) versus line- (vector) based approaches to quantifying measurements (Thewlis et al. 2019) and which method minimises variance, while still adequately quantifying the trait in question.

Results of the workshop: Both subgroups were able to establish working pipelines to assess their training datasets (which they brought with them).

The image subgroup was interested in generating landmarks and extracting linear measurements for teeth (B. Wynd, C. Charpentier, B. Khakurel). They were able to generate measurements on a training dataset of 50 images (both landmark and linear measurements) and then used ML-morph (Porto and Voje 2020) to train the ML-morph machine-learning model to generate measurements for a full dataset of ~ 150 images. An overview of the pipeline (Fig. 1) is available at the workshop's GitHub repository.*[11] Importantly, immediate results indicate that the training dataset was too small and the estimated measurements were, thus, not reliable. However, the pipeline is fully functional and only requires further manipulation by the image subgroup to evaluate necessary training sample sizes to provide confident quantification of images.

The video subgroup was able to generate video annotations using DeepLabCut (Nath et al. 2019) for different parts of butterflies while feeding on a flower (C. Lawrence). This subgroup started by hand-annotating a few videos collected from the Smithsonian Tropical Research Institute. These annotations were then used to train a model. The model performed well and managed to annotate the desired body parts on multiple individuals. These videos were then passed to OpenCV (Bradski 2000), an open source computer vision software library, to detect an ArUco tag (i.e. an identifier for each individual in a multi-animal setup) that had been placed on the underside of the butterfly's wing. While the annotations resolved well across different butterfly videos, the ArUco tag was not detected. However, the attempts to: (i) annotate videos of butterfly behaviour using DeepLabCut and (ii) place ArUco tags on butterflies and track them, were the first of their kind.

Another effort this subgroup undertook was sifting through courtship behaviour data collected on jumping spiders (L. Yan). Key body parts of male spiders were labelled using DeepLabCut as well and fixed body parts were used for landmarks for procrustes analysis (a shape analysis that accounts for and unifies the same object appearing at different angles, size and angles in images/videos while removing the effect of size) to

account for variations in spider shape and video variations. Having quantified the different types of spider behaviour into a multi-dimensional space, based on posture coordinates and primary motion measurements, this subgroup sought a way to code the space by stereotypical behavioural units. To do this, they reduced the dimensionality of the data to two dimensions for easier visualisation and for performing clustering algorithms. Several clustering algorithms (e.g. k-means (Hartigan and Wong (1979)), hierarchical clustering (Johnson 1967)) were performed to group similar types of behaviour together and separate distinct types. They used a watershed algorithm in *patternize* (an R programming package; Van Belleghem et al. (2017)) to draw boundaries around distinct groups of behaviour (clusters). While they were able to generate clusters and segment basic units, interpreting the results was not possible before the end of the workshop.

Future directions/plans/recommendations: The data quantification project group has plans to continue annotating their biological data (images and videos) for use as training datasets. The image subgroup will be looking to publish a manuscript focusing on the pipeline and requirements for the training dataset, best practices in automating measurements and an assessment of the variance in linear- versus landmark-based measurements. The present goal is to publish a manuscript led by (and part of the dissertation of a graduate student) C. Charpentier. This manuscript will be a small step forward in the application of machine-learning to expedite the data collection process in landmark-based analyses of image data. The video subgroup will further look into details in quality control of annotated landmarks, especially inconsistencies between frames, to generate more accurate annotations that capture the differences between types of behaviour instead of filming methods (e.g. noisy background, individual size and angle differences). Additionally, it is important to develop a measure to account for consistent ArUco tag tracking with the presence of attenuation of the tag when the video is filmed at varied angles. The video group worked mostly with videos taken from natural or inconsistent backgrounds, representing the majority of animal behaviour recordings. Despite its prevalence, it is more challenging for automatic quantification than analysing model species and uniform background. The pipeline is aimed to be broadly applicable to naturalistic video-data quantification and set the stage for higher resolution of behaviour categorisation.

Project III: A Graph Approach to Understand Complexity in Species

Problem to solve: Complexity in living systems is an elusive concept usually defined in terms of a raw number of constituent elements, how they connect to each other and the number of hierarchical levels in which those elements can be organised. One alternative technique to describe such systems is to use graphs with anatomical entities represented as nodes and the relationships between them as edges. The goal is to use these constructed graphs as an approach to assess complexity across the tree of life. Over the course of the workshop, the group (D. Sasso Porto, E. Alhajjar, H. Lapp and J. N. Keating) developed a pilot pipeline for achieving such a task.

Approach: To develop and test our pipeline, this group retrieved the phylogenetic character matrix from Mirande (2018)—a morphological study of fishes in the family Characidae—as available in the Phenoscape Knowledgebase (KB) using the *rphenoscape* R package (Lapp et al. 2022). The data from the Phenoscape KB comprises semantic phenotypes; i.e. phylogenetic characters in which anatomical entities and their qualities are annotated with ontology terms. All anatomical entities connected by *part_of* relations were extracted from the semantic phenotypes and employed to build a graph of dependencies (Fig. 2A) using the *igraph* R package (Csardi and Nepusz 2006). The graph was polarised by the addition of a 'root' vertex: an invariable entity from which all other entities are dependent (e.g. multicellular organism). Indirect dependencies were removed using the remove_indirect() function from the rphenoscape package (Lapp et al. 2022). The group used four graph-based metrics to characterise the complexity of each subgraph representing the anatomy of an individual species. These were: number of vertices (the number of anatomical entities present in a phenotype), number of edges (the number of direct dependencies in a phenotype), mean path distance to root vertex (the mean number of nested dependencies) and edge density (the ratio of the number of edges to the maximum possible number of edges, a measure of the integration of anatomical entities by dependencies). These metrics were used as data for each extant species to evaluate different ways to reconstruct the evolution of 'complexity' using standard phylogenetic comparative methods, such as those implemented in the *phytools* R package (Revell 2011). For a proof-of-concept to illustrate the proposed framework, we employed a simple ancestral state reconstruction using a topology extracted from the *fishtree* package (Chang et al. 2019), analysed under Brownian motion (Fig. 2B), as implemented by the function *contmap**[38] from *phytools*.

Results of the workshop: We created a pilot pipeline, PhenoNet, to study the evolution of complexity. The pipeline is available as an R script in the TraitFest2023 repository.*[11] Using this new tool, we were able to create many graphs to validate our assumptions and test the new tool.

Future directions/plans/recommendations: The pipeline developed in the workshop used semantic data from the Phenoscape KB as the pilot study; this pipeline can be applied to any dataset for which anatomical entities can be annotated with ontology terms from an anatomy ontology (e.g. Uberon;*[12] Mungall et al. (2012)), Hymenoptera Anatomy Ontology (HAO;*[13] Yoder et al. (2010)), Ontology for the Anatomy of Insect SkeletoMuscular System (AISM;*[10] Girón et al. (2023)). Graph-based representations of organismal anatomy open new opportunities for exploring alternative metrics to assess not only complexity, but also similarity amongst phenotypes of organisms.

Projects IV, V & VI: Trait Repository: Creating a "GenBank" for Phenotypic Data

Problem to solve: Numerous individuals and projects display a keen interest in the study of phenotypes, encompassing morphological, anatomical, ecological and physiological characteristics. These traits play a crucial role in various fields, such as phylogenetic, evo-devo, population and ecological research. The scientific community has a few,

emergent resources dedicated to trait data, for example, the Functional Trait Resource for Environmental Studies (FuTRES) datastore (Balk et al. 2022) and Open Traits Network (Gallagher et al. 2020). Open Traits Network is akin to Dryad, registering trait databases and datasets; FuTRES is akin to GenBank®, which is a searchable datastore for trait data measured on invidividual organisms or specimens and where trait measurements are defined by anatomical terms structured in an ontology. FuTRES interacts with other repositories like the Global Biodiversity Information Facility (GBIF) and OpenContext and currently only accepts linear trait data, leaving a need for integration of other types of trait data (e.g. ordinal, character, behavioural) and at different levels of biotic organisation (e.g. population).

The ideal repository would possess the following key features:

- **Inclusivity:** It should readily accept a diverse range of traits, including phenotypic, ecological and distributional data, from all taxonomic groups.
- **User-friendly query service:** The repository should offer a user-friendly interface that allows researchers to effortlessly query and retrieve essential phenotypic information for a given taxon or specimen.
- **Interoperability:** To enhance collaboration and data integration, the repository must be compatible with existing platforms like GenBank and GBIF, ensuring seamless interaction with other essential repositories.

Approach: Our group was composed of ontologists (J. Balhoff, M.A. Balk, P. Mabee), ontology-oriented entomologists (J. Girón, I. Mikó, G. Montanaro, M. Rossini, K. C. Seltmann, S. Tarasov) and ecologists (I. Fluck, A. Espindola). To address this issue, we developed a prototype repository named PhenoRepo*[14] and populated it with relevant use cases.

In order to assess PhenoRepo's functionality, we chose to focus on two insect groups: Coleoptera (beetles) and Hymenoptera (bees, ants and wasps). The Coleoptera dataset mainly consisted of species of dung beetles, while the Hymenoptera dataset included various bee species, such as *Agapostemon texanus* Cresson, 1872 (family Halictidae),*[15] *Chalepogenus caeruleus* (Friese, 1906) (family Apidae),*[16] as well as species from the wasp genera *Gryonoides* (family Scelionidae)*[17] and *Cephus* (family Cephidae).*[18]

Bees were chosen as an exemplary group due to their crucial role as pollinators and the concerning global decline they are facing (Koh et al. 2015). Consequently, ecologists and evolutionary biologists prioritise their study. However, the underlying reasons for declines specific to certain bee taxa or regions remain poorly understood, hindering effective conservation efforts (Menz et al. 2011). Although substantial data on bees exist (Seltmann et al. 2021), they are scattered across various formats and lack a unified means of translation into semantic statements or computable repositories until PhenoRepo's development.

PhenoRepo Design. Using a semantic approach to describe phenotypes (employing terms sourced from relevant ontologies) has proven to be a potent tool for rendering

phenotypes understandable and accessible to computers (Balhoff et al. 2010, Balhoff et al. 2013, Balhoff et al. 2014, Vogt 2019). PhenoRepo adopts this approach to great effect and is modelled after the FuTRES ontological backbone and triple store (see Balk et al. (2022)). An elementary piece of data (= phenotypic observation) in PhenoRepo represents an ontology individual (i.e. an instance or concrete object) that may be linked with other individuals via object, data or annotation properties. In this way, phenotypes are expressed as knowledge graphs and PhenoRepo exhibits a set of these knowledge graphs in the form of OWL (Web Ontology Language)*[19] files that may be submitted by users. OWL is the standard file format for ontologies. PhenoRepo OWL files are converted to RDF (Resource Description Framework)*[20] format and represent a triplestore that can be queried using SPARQL (a query language for RDF).*[21] Users may download the query output, which represents particular traits for certain taxa, to their computers for further analysis. In order to facilitate prototyping, PhenoRepo is currently available as a GitHub repository, where authorised users can upload data using standard GitHub workflows.

Submitting Data to PhenoRepo. To contribute trait data to PhenoRepo, users are required to represent their data semantically in the form of knowledge graphs. While these graphs can be constructed using the widely-used Protégé software, it may not be the most straightforward approach. Instead, we recommend using specialised software designed explicitly for this purpose, as follows:

- **Phenex:***[8] Users should employ Phenex to describe morphological traits in the format of character matrices (Balhoff et al. 2010, Balhoff et al. 2014).
- **Phenoscript:***[9] For morphological or ecological traits, written as character statements, Phenoscript is the preferred software to use (Mikó et al. 2021).
- **ROBOT templates:***[22] When dealing with custom, medium-sized phenotypes, users can rely on ROBOT templates to facilitate the conversion into knowledge graphs (Jackson et al. 2019).

By utilising these specialised tools, users can effectively construct semantic knowledge graphs, making the process of submitting trait data to PhenoRepo more efficient and seamless.

Results of the workshop: We created the instance-based repository PhenoRepo, a repository of semantic traits for any phenotype, including morphological, ecological and environmental data. Users were able to upload OWL files to the PhenoRepo GitHub repository.

PhenoRepo was tested with data from diverse sources, including two Darwin Core-formatted (Darwin Core Task Group 2009, Wieczorek et al. 2012) bee datasets. Specifically, we mapped Darwin Core specimen occurrence data to existing ontologies and converted the comma-delimited files to Phenoscript. Once in Phenoscript, the statements were further converted into OWL files and uploaded to the PhenoRepo for indexing into the repository. To achieve this objective, the Darwin Core catalogNumber*[23] had to be included in Phenoscript syntax. The Darwin Core catalogNumber serves as an

identifier for specimens within a dataset and its inclusion allows for referencing specimen-specific information within Phenoscript.

Additionally, we wrote a workflow and infrastructure to apply reasoning to the data. Environmental data and measurement data were also converted into Phenoscript format, which then was converted to OWL and uploaded to PhenoRepo. These Phenoscript files were also converted to human-readable Markdown syntax, which may be included in journal publications. We also managed to convert character matrices annotated in Phenex into PhenoRepo by using a custom Python script. Example input and output files can be found on the PhenoRepo and in our workshop presentation.*[24]

Future directions/plans/recommendations: There is a significant need to expand our ability to include trait and phenotype data within a semantic framework. To do so, ontology resources need to be expanded and additional effort aligning existing ontologies is needed to make inferences across taxa. Already, M.A. Balk is working with the Ontology of Biological Attributes (OBA; Stefancsik et al. (2023)) and the Uber Anatomy Ontology (Uberon; Mungall et al. (2012)) to add trait terms. One of the major constraints that remains in producing semantic descriptions is the lack of standard logical models for describing specific types of phenotypes. As a result, rapid phenotypic descriptions are hindered, a problem that can only be resolved if the entire community is involved in the development of such models and standards.

Indeed, PhenoRepo was a successful proof of concept, demonstrating a broader approach to taxon-independent phenotype database and workflow; however, more effort is needed to describe functional traits (Violle et al. 2007), map them to existing ontologies or create novel ontologies. The trait repository working group has plans to continue creating use cases for traits, including bee traits, which can provide additional use cases for Phenoscript. In the short term, group members held a symposium at the 10th Congress of International Society of Hymenopterists*[25] and to include Phenoscript uploaded to PhenoRepo in a manuscript describing bee functional traits.

Project VII: Image extraction from literature

Problem to solve: Images are a useful tool in the biological sciences to convey visual information about organisms, including anatomical features and contrasting differences between species. The scientific literature contains many potentially useful images of organisms and, often, an accompanying caption with relevant textual information, such as a scale or trait descriptions. The ability to extract images and their accompanying captions can potentially greatly increase the amount of trait information accessible to researchers through resources, such as the Phenoscape KB. Our goal is to create a workflow to extract images and captions from PDFs into usable formats that can be used downstream to mine captions for anatomical terms and automatically annotate traits from images by using a machine-learning pipeline like ML-morph (Porto and Voje 2020) or ChatGPT.*[26] The group focused their efforts on the literature pertaining to Bryozoa, the phylum of colonial aquatic invertebrates composed of modular units called zooids, as both A. Porto and M.A. Balk have the resources and interest in this group.

Approach: The group included participants with expertise or interest in ontologies and Phenoscape (M.A. Balk, W. Dahdul, J. Balhoff, D. Sasso Porto), a machine-learning expert (A. Porto) and persons interested in ChatGPT (A. Porto, J. Balhoff).

Both goals of this project required the extraction of images and captions from a collection of PDFs pertaining to bryozoans. A literature search previously compiled by A. Porto was used as the image and text corpus. As a test run, the most recent literature from 2016–2021 (441 PDFs) was used as these papers were more likely to have a standardised format and are computer-readable with Optical Character Recognition text. To extract images and captions from PDFs, they used PDFFigures 2.0, a Scala-based tool developed by AllenAI.*[28]

The first goal was to extract trait terms from the textual descriptions in captions. This goal necessitated the creation of a new ontology for bryozoan anatomy and traits. The terms and species identities will feed into the Phenoscape KB to create character matrices. For the extraction of visual traits from images, we planned to utilise the machine-learning tools developed by A. Porto to automatically segment and landmark images of bryozoans (see Porto and Voje (2020), Di Martino et al. (2022)).

Results of the workshop:

The PDFFigures 2.0 application successfully extracted 2874 images in JPG format from the 441 PDF files. As the output images may be of different types (e.g. Scanning Electron Microscope (SEM) images, tables, maps), they performed a Principal Component Analysis (PCA) to group the images by type. The scripts and a demo are on the group's repository, lit-bryo.*[27] The PCA revealed that the programme did a surprisingly good job of recognising non-SEM images of bryozoans and SEM images not of bryozoans.

M.A. Balk created the Bryozoan Attribute Ontology (BAO)*[29] using the Ontology Development Kit (Matentzoglu et al. 2022) building off terms in Uberon (Mungall et al. 2012) and OBA (Ontology of Biological Attributes;*[30] Stefancsik et al. (2023)) ontologies. BAO contains 10 new terms created during the workshop related to morphological structures present on bryozoans (Table 1).

Future directions/plans/recommendations: The group's future plan is to create a complete workflow from image extraction to term generation and morphology analyses (Fig. 3). The next steps in realising this pipeline are to use Porto's ML pipeline to annotate traits from the extracted images and to parse their associated figure legends to capture relevant information about traits.

The group is also continuing to develop the BAO, with feedback from the OBO Foundry community (Smith et al. 2007), where the ontology will eventually be published. We plan to add new terms as needed for representing traits related to those extracted from the images and captions.

Project VIII: ubeRsim **: an R package to implement semantic similarity methods for pairwise and profile similarity**

Problem to solve: RPhenoscape*[33] is an R package providing convenient access to the data, ontologies and semantics-based analytics offered by the Phenoscape Knowledgebase (KB). In particular, it includes methods for computing a number of widely-used semantic similarity metrics, both between pairs of ontology terms ("pairwise similarity") and between groups of terms ("profile similarity") (Pesquita et al. 2009), with different ways of aggregating the pairwise scores per group. These metrics form a key part of enabling traits to be fully computable, including for comparative phylogenetic analysis (Porto et al. 2022). However, this necessarily requires that the respective ontology is included in the Phenoscape KB and, thus, in its build pipeline. Extending this pipeline is a non-trivial undertaking, even for its maintenance team and is challenging to make responsive to the needs of researchers. In contrast, the Ubergraph*[34] project has created an RDF database with a public SPARQL query endpoint, an interoperable graph data dump and a reproducible build pipeline for the large collection of ontologies under the OBO Library umbrella (Balhoff et al. 2022). The build pipeline ensures that all OWL entailments (inferences implied by, but not expressly asserted in the source ontologies) from each ontology and by the integration of ontologies are included in the RDF database through pre-reasoning and materialising the inferred axioms. Our goal was to demonstrate that the computational semantics capabilities available from within the RPhenoscape package can be implemented on top of an Ubergraph RDF database, focusing specifically on the semantic similarity metrics. If successful, this would enable researchers to use these metrics for any ontology within the OBO Library collection, eliminating a major hurdle to adoption of these capabilities beyond the domains (such as vertebrate anatomy) currently encompassed by the Phenoscape KB.

Approach: The group (H. Lapp, lead developer of RPhenoscape and J. Balhoff, lead developer of the Ubergraph RDF database) set out to create an R package that would re-implement the semantic similarity methods from the RPhenoscape package for both pairwise and profile similarity, by querying the public Ubergraph RDF database instance through its SPARQL endpoint.*[31]

Results of the workshop: The group created a working proof-of-concept in the form of an R package tentatively called ubeRsim.*[35] The implemented methods allow calculating graph-based (a.k.a. edge-based) semantic similarity metrics, including Jaccard*[36] and Cosine*[37] similarity, both for pairwise and for profile similarity. Information content-based semantic similarity scores, such as Resnik similarity (Resnik 1995, Resnik 1999, Lord et al. 2003), require for each term the probability of encountering it, which is normally derived from the frequencies of terms in a chosen text corpus. As a resource agnostic to specific research questions and, thus, the appropriate choice of corpus, term frequencies at least at this time are not available from Ubergraph.

The algorithm implemented in RPhenoscape for calculating semantic similarity scores uses matrix multiplication, which is very efficient in R, but requires a so-called subsumer matrix $(M_{ij} = \begin{cases} 1 \text{ if } T_i \sqsupseteq T_j \\ 0 \text{ otherwise} \end{cases})$. An endpoint in the Phenoscape KB API, which RPhenoscape queries, assembles and returns this matrix from a given list of input terms. In contrast, a SPARQL query of an RDF graph (as well as an SQL query of an equivalent table of graph

edges) can only return an adjacency list. ubeRsim converts this to a subsumer matrix to enable keeping the same efficient matrix multiplication-based algorithm for computing semantic similarity scores.

They compared Jaccard similarity scores obtained through the ubeRsim implementation (and, thus, from subsumption subgraphs obtained from the public Ubergraph instance) with those returned from RPhenoscape for a list of select Uberon ontology terms (vertebrate limbs and paired and unpaired fins). They found that although the similarity scores were not numerically identical, they were similar and their relative order was mostly the same. These differences can, in part, be traced back to the upper ontologies included in the pre-reasoning, which differs between the two underlying databases and in the difference of anonymous OWL class expressions (such as "part_of some X", where X is a named term in an ontology) that are materialised in the Phenoscape KB, but are not in Ubergraph as a broader-purpose resource.

Future directions/plans/recommendations: The major directions for taking ubeRsim from its current proof-of-concept stage to a more widely-usable package in the comparative trait analysis ecosystem in R include the following: (1) generating "virtual" subsumer terms equivalent to anonymous OWL class expressions that would have been encountered as subsumers had they been materialised, so as to achieve equally discriminatory similarity scores as RPhenoscape and the Phenoscape KB; (2) providing user choice for which properties beyond subclass relationships to use for querying subsumption subgraphs; and (3) finding a mechanism or resource for obtaining broadly applicable term frequencies to enable information content-based similarity metrics.

Postworkshop Survey Responses

A post-workshop survey was sent to participants to understand their experience and evaluate the potential impact of the workshop. We received 10 highly positive responses to questions about whether the workshop was worth their time (4.9 average response on a 1-5 scale), whether they gained new knowledge (4.9) and whether they made connections with other participants that will enable new or better research endeavours (4.8). In free-form responses, participants also responded positively to the OST format and valued the unstructured work time and learning opportunities. On the other hand, a desire for more time in the initial project exploration phase, scheduled bootcamps so that everyone has an opportunity to attend and the desire for more social gatherings and meals together was noted.

Discussion

Although the Phenoscape KB is broad in scope by including Uberon*[12] as the backbone cross-species anatomy ontology for animals, its taxonomic focus is on vertebrates. One major goal of the workshop was to find points of convergence where tools and pipelines can be extended and made applicable across different taxonomic groups, such as invertebrates. For this reason, the target participants for the workshop

were people with broadly diverse interests and backgrounds. This was certainly what made the workshop very unique, as well as challenging in some instances.

The use of Open Space Technology (OST): the good and the not so great. Giving people the freedom to choose which projects to work on and self-organise significantly distributed the weight of coordinating activities, which streamlined the flow of ideas by allowing groups to focus on particular tasks and people to shift to other projects when it seemed appropriate. By setting up the workshop wiki,*[4] the list of participants*[2] with their interests and background and the list of potential projects (as issues in the GitHub repository*[32]), we tried to get people to identify ideas and potential partners to work with, before getting together. Still, most of the self-organising took place during the workshop and that may have reduced the amount of time available for groups to work together.

Another aspect of the freedom offered by the OST format was that by working together in one place, people felt that there was time and space to get things done, even though the time was limited. For example, members within and between groups interacted with one another frequently, progressing projects along more quickly than if we were working asynchronously.

A diversity of topics for a diversity of users and backgrounds. During the introductions session, it seemed unclear how we would manage to get people with such a broad assortment of backgrounds to work together on innovating uses for the tools provided by the Phenoscape KB. Fortunately, participants came to the workshop willing to learn and share. The OST also allowed the integration of ideas originating from the different approaches to data that dissimilar expertises can bring and how the ideas became more solid by working together with people with varied perspectives.

Bootcamps: necessary distractions? The improptu bootcamps developed during the workshop made it so that all the participants went home with new knowledge besides the products of the developed projects, but, for some groups, the bootcamps distracted group members from contributing to achieving the goals of their projects. The bootcamps also empowered participants from different backgrounds to learn and understand what skills people had and have a better idea about how to integrate them.

Extending beyond our usual networks. Perhaps one of the greatest achievements of the event was exposing researchers who are experts in a particular domain to explore and learn from experts with very different backgrounds from their own. As researchers, we tend to attend the same kinds of events on a repeating basis, talking primarily with those in our own field. Bringing people from diverse backgrounds together and encouraging them to work in interdisciplinary groups resulted in a unique opportunity for professional development, especially for those who are in the earlier stages of their careers. The workshop allowed everyone to be the expert and be introduced to a completely new subject.

Conclusions

As we believe the results and observations we report here show, developing interdisciplinary meetings can be extremely productive, with both tangible and intangible outcomes greatly outweighing the organisational costs, even if these are substantial and especially so for projects that aim to integrate across knowledge domains.

Bringing people together with complementary knowledge, skills and interests and getting them to talk to each other and teach each other towards shared research objectives, is a powerful tool to expand perspectives and helps foster a sense of community and belonging.

Grant title

Collaborative Research: ABI Innovation: Enabling machine-actionable semantics for comparative analyses of trait evolution

Awards 1661456 (Duke University), 1661529 (University of South Dakota), 1661516 (Virginia Tech), and 1661356 (University of North Carolina at Chapel Hill and RENCI). The attendance of S. Tarasov and D. S Porto was supported by the Research Council of Finland (#339576 and #346294).

Hosting institution

Duke University, University of South Dakota, Virginia Tech, University of North Carolina at Chapel Hill, and RENCI.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Balhoff J, Dahdul W, Kothari C, Lapp H, Lundberg J, Mabee P, Midford P, Westerfield M, Vision T (2010) Phenex: Ontological Annotation of Phenotypic Diversity. PLoS ONE 5 (5). https://doi.org/10.1371/journal.pone.0010500
- Balhoff J, Mikó I, Yoder M, Mullins P, Deans A (2013) A Semantic Model for Species Description Applied to the Ensign Wasps (Hymenoptera: Evaniidae) of New Caledonia. Systematic Biology 62 (5): 639-659. https://doi.org/10.1093/sysbio/syt028
- Balhoff J, Dahdul WM, Dececchi T, Lapp H, Mabee PM, Vision TJ (2014) Annotation of phenotypic diversity: decoupling data curation and ontology curation using Phenex. Journal of Biomedical Semantics 5 (1). https://doi.org/10.1186/2041-1480-5-45

- Balhoff J, Phenoscape Project Team (2016) The Phenoscape Knowledgebase: tools and APIs for computing across phenotypes from evolutionary diversity and model organisms. bioRxiv https://doi.org/10.1101/071951
- Balhoff J, Bayindir U, Caron AR, Matentzoglu N, Osumi-Sutherland D, Mungall CJ (2022) Ubergraph: integrating OBO ontologies into a unified semantic graph. 1613. ICBO-2022: International Conference on Biomedical Ontology. URL: https://icbo-conference.github.io/icbo2022/papers/ICBO-2022_paper_5005.pdf
- Balk M, Deck J, Emery K, Walls R, Reuter D, LaFrance R, Arroyo-Cabrales J, Barrett P, Blois J, Boileau A, Brenskelle L, Cannarozzi N, Cruz JA, Dávalos L, de la Sancha N, Gyawali P, Hantak M, Hopkins S, Kohli B, King J, Koo M, Lawing AM, Machado H, McCrane S, McLean B, Morgan M, Pilaar Birch S, Reed D, Reitz E, Sewnath N, Upham N, Villaseñor A, Yohe L, Davis E, Guralnick R (2022) A solution to the challenges of interdisciplinary aggregation and use of specimen-level trait data. iScience 25 (10). https://doi.org/10.1016/j.isci.2022.105101
- Bradski G (2000) The openCV library. Dr. Dobb's Journal: Software Tools for the Professional Programmer 25 (11): 120-123.
- Chang J, Rabosky D, Smith S, Alfaro M (2019) An R package and online resource for macroevolutionary studies using the ray-finned fish tree of life. Methods in Ecology and Evolution 10 (7): 1118-1124. https://doi.org/10.1111/2041-210x.13182
- Csardi G, Nepusz T (2006) The igraph software package for complex network research. 0.10.4. Release date: 2023-1-27. URL: https://igraph.org
- Dahdul W, Balhoff J, Engeman J, Grande T, Hilton E, Kothari C, Lapp H, Lundberg J, Midford P, Vision T, Westerfield M, Mabee P (2010) Evolutionary Characters, Phenotypes and Ontologies: Curating Data from the Systematic Biology Literature. PLoS ONE 5 (5). https://doi.org/10.1371/journal.pone.0010708
- Dahdul W, Dececchi TA, Ibrahim N, Lapp H, Mabee P (2015) Moving the mountain: analysis of the effort required to transform comparative anatomy into computable anatomy. Database 2015 https://doi.org/10.1093/database/bav040
- Darwin Core Task Group (2009) Darwin Core. Biodiversity Information Standards (TDWG). http://www.tdwg.org/standards/450. Accessed on: 2023-9-06.
- Dececchi TA, Balhoff J, Lapp H, Mabee P (2015) Toward Synthesizing Our Knowledge of Morphology: Using Ontologies and Machine Reasoning to Extract Presence/Absence Evolutionary Phenotypes across Studies. Systematic Biology 64 (6): 936-952. https://doi.org/10.1093/sysbio/syv031
- Di Martino E, Berning B, Gordon DP, Kuklinski P, Liow LH, Ramsfjell MH, Ribeiro HL, Smith AM, Taylor PD, Voje KL, Waeschenbach A, Porto A (2022) DeepBryo: a web app for AI-assisted morphometric characterization of cheilostome bryozoans. bioRxiv https://doi.org/10.1101/2022.11.17.516938
- Edmunds R, Su B, Balhoff J, Eames BF, Dahdul W, Lapp H, Lundberg J, Vision T, Dunham R, Mabee P, Westerfield M (2015) Phenoscape: Identifying Candidate Genes for Evolutionary Phenotypes. Molecular Biology and Evolution 33 (1): 13-24. https://doi.org/10.1093/molbev/msv223
- Fernando P, Mabee P, Zeng E (2023) Protein–protein interaction network module changes associated with the vertebrate fin-to-limb transition. Scientific Reports 13 (1). https://doi.org/10.1038/s41598-023-50050-2
- Gallagher R, Falster D, Maitner B, Salguero-Gómez R, Vandvik V, Pearse W, Schneider F, Kattge J, Poelen J, Madin J, Ankenbrand M, Penone C, Feng X, Adams V, Alroy J,

Andrew S, Balk M, Bland L, Boyle B, Bravo-Avila C, Brennan I, Carthey AR, Catullo R, Cavazos B, Conde D, Chown S, Fadrique B, Gibb H, Halbritter A, Hammock J, Hogan JA, Holewa H, Hope M, Iversen C, Jochum M, Kearney M, Keller A, Mabee P, Manning P, McCormack L, Michaletz S, Park D, Perez T, Pineda-Munoz S, Ray C, Rossetto M, Sauquet H, Sparrow B, Spasojevic M, Telford R, Tobias J, Violle C, Walls R, Weiss KB, Westoby M, Wright I, Enquist B (2020) Publisher Correction: Open Science principles for accelerating trait-based science across the Tree of Life. Nature Ecology & Evolution 4 (4): 662-662. https://doi.org/10.1038/s41559-020-1167-9

- Girón JC, Tarasov S, González Montaña LA, Matentzoglu N, Smith AD, Koch M, Boudinot BE, Bouchard P, Burks R, Vogt L, Yoder M, Osumi-Sutherland D, Friedrich F, Beutel R, Mikó I (2023) Formalizing Invertebrate Morphological Data: A Descriptive Model for Cuticle-Based Skeleto-Muscular Systems, an Ontology for Insect Anatomy, and their Potential Applications in Biodiversity Research and Informatics. Systematic Biology https://doi.org/10.1093/sysbio/syad025
- Glimm B, Horrocks I, Motik B, Stoilos G, Wang Z (2014) HermiT: An OWL 2 Reasoner. Journal of Automated Reasoning 53 (3): 245-269. https://doi.org/10.1007/s10817-014-9305-1
- Haendel MA, Balhoff JP, Bastian FB, Blackburn DC, Blake JA, Bradford Y, Comte A, Dahdul WM, Dececchi TA, Druzinsky RE, Hayamizu TF, Ibrahim N, Lewis SE, Mabee PM, Niknejad A, Robinson-Rechavi M, Sereno PC, Mungall CJ (2014) Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. Journal of Biomedical Semantics 5 (1). https://doi.org/10.1186/2041-1480-5-21
- Hartigan JA, Wong MA (1979) Algorithm AS 136: A K-Means Clustering Algorithm. Applied Statistics 28 (1). https://doi.org/10.2307/2346830
- Jackson R, Balhoff J, Douglass E, Harris N, Mungall C, Overton J (2019) ROBOT: A Tool for Automating Ontology Workflows. BMC Bioinformatics 20 (1). https://doi.org/10.1186/s12859-019-3002-3
- Johnson S (1967) Hierarchical clustering schemes. Psychometrika 32 (3): 241-254. https://doi.org/10.1007/bf02289588
- Kazakov Y, Krötzsch M, Simančík F (2013) The Incredible ELK. Journal of Automated Reasoning 53 (1): 1-61. https://doi.org/10.1007/s10817-013-9296-3
- Koh I, Lonsdorf E, Williams N, Brittain C, Isaacs R, Gibbs J, Ricketts T (2015) Modeling the status, trends, and impacts of wild bee abundance in the United States. Proceedings of the National Academy of Sciences 113 (1): 140-145. https://doi.org/10.1073/pnas.1517685113
- Lapp H, Xu H, Bradley J (2022) rphenoscape: Semantically Rich Phenotypic Traits from the Phenoscape Knowledgebase. URL: http://rphenoscape.phenoscape.org
- Lord PW, Stevens RD, Brass A, Goble CA (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19 (10): 1275-1283. https://doi.org/10.1093/bioinformatics/btg153
- Mabee P, Balhoff JP, Dahdul WM, Lapp H, Midford PE, Vision TJ, Westerfield M (2012) 500,000 fish phenotypes: The new informatics landscape for evolutionary and developmental biology of the vertebrate skeleton. Journal of Applied Ichthyology 28 (3): 300-305. https://doi.org/10.1111/j.1439-0426.2012.01985.x
- Mabee PM, Dahdul WM, Balhoff JP, Lapp H, Manda P, Uyeda J, Vision TJ, Westerfield M (2018) Phenoscape: semantic analysis of organismal traits and genes yields insights in evolutionary biology. In: Thessen A (Ed.) Application of semantic technology in

biodiversity science. IOS Press, Berlin, 207-224 pp. https://doi.org/10.3233/978-1-61499-854-9-207
- Manda P, Balhoff J, Lapp H, Mabee P, Vision T (2015) Using the phenoscape knowledge base to relate genetic perturbations to phenotypic evolution. genesis 53 (8): 561-571. https://doi.org/10.1002/dvg.22878
- Marshall CR, Finnegan S, Clites EC, Holroyd PA, Bonuso N, Cortez C, Davis E, Dietl GP, Druckenmiller PS, Eng RC, Garcia C, Estes-Smargiassi K, Hendy A, Hollis KA, Little H, Nesbitt EA, Roopnarine P, Skibinski L, Vendetti J, White LD (2018) Quantifying the dark data in museum fossil collections as palaeontology undergoes a second digital revolution. Biology Letters 14 (9). https://doi.org/10.1098/rsbl.2018.0431
- Matentzoglu N, Goutte-Gattat D, Tan SZK, Balhoff JP, Carbon S, Caron AR, Duncan WD, Flack JE, Haendel M, Harris NL, Hogan WR, Hoyt CT, Jackson RC, Kim H, Kir H, Larralde M, McMurry JA, Overton JA, Peters B, Pilgrim C, Stefancsik R, Robb SM, Toro S, Vasilevsky NA, Walls R, Mungall CJ, Osumi-Sutherland D (2022) Ontology Development Kit: a toolkit for building, maintaining and standardizing biomedical ontologies. Database 2022 https://doi.org/10.1093/database/baac087
- Menz MM, Phillips R, Winfree R, Kremen C, Aizen M, Johnson S, Dixon K (2011) Reconnecting plants and pollinators: challenges in the restoration of pollination mutualisms. Trends in Plant Science 16 (1): 4-12. https://doi.org/10.1016/j.tplants.2010.09.006
- Mikó I, Masner L, Ulmer J, Raymond M, Hobbie J, Tarasov S, Margaría CB, Seltmann K, Talamas E (2021) A semantically enriched taxonomic revision of *Gryonoides* Dodd, 1920 (Hymenoptera, Scelionidae), with a review of the hosts of Teleasinae. Journal of Hymenoptera Research 87: 523-573. https://doi.org/10.3897/jhr.87.72931
- Mirande JM (2018) Morphology, molecules and the phylogeny of Characidae (Teleostei, Characiformes). Cladistics 35 (3): 282-300. https://doi.org/10.1111/cla.12345
- Mungall C, Gkoutos G, Washington N, Lewis S (2007) Representing Phenotypes in OWL. In: Golbreich C, Kalyanpur A, Parsia B (Eds) OWL: Experiences and Directions. Innsbruck, Austria. URL: https://ceur-ws.org/Vol-258/paper29.pdf
- Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA (2012) Uberon, an integrative multi-species anatomy ontology. Genome Biology 13 (1). https://doi.org/10.1186/gb-2012-13-1-r5
- Nath T, Mathis A, Chen AC, Patel A, Bethge M, Mathis MW (2019) Using DeepLabCut for 3D markerless pose estimation across species and behaviors. Nature Protocols 14 (7): 2152-2176. https://doi.org/10.1038/s41596-019-0176-0
- Owen H (2008) Open space technology: A user's guide. Berrett-Koehler Publishers, 192 pp.
- Pesquita C, Faria D, Falcão A, Lord P, Couto F (2009) Semantic Similarity in Biomedical Ontologies. PLoS Computational Biology 5 (7). https://doi.org/10.1371/journal.pcbi.1000443
- Porto A, Voje K (2020) ML-morph: A fast, accurate and general approach for automated detection and landmarking of biological structures in images. Methods in Ecology and Evolution 11 (4): 500-512. https://doi.org/10.1111/2041-210x.13373
- Porto D, Tarasov S, Charpentier C, Lapp H, Balhoff J, Vision T, Dahdul W, Mabee P, Uyeda J (2023) rphenoscate: An R package for semantic-aware evolutionary analyses of anatomical traits. Methods in Ecology and Evolution 00: 1-10. https://doi.org/10.1111/2041-210X.14210

- Porto DS, Dahdul WM, Lapp H, Balhoff JP, Vision TJ, Mabee PM, Uyeda J (2022) Assessing Bayesian Phylogenetic Information Content of Morphological Data Using Knowledge From Anatomy Ontologies. Systematic Biology 71 (6): 1290-1306. https://doi.org/10.1093/sysbio/syac022
- Resnik P (1995) Using Information Content to Evaluate Semantic Similarity in a Taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1. URL: http://dl.acm.org/citation.cfm?id=1625855.1625914 [ISBN 978-1-55860-363-9].
- Resnik P (1999) Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. J. Artif. Intell. Res. 11 (95). URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.50.3785
- Revell L (2011) phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3 (2): 217-223. https://doi.org/10.1111/j.2041-210x.2011.00169.x
- Seltmann K, Allen J, Brown B, Carper A, Engel M, Franz N, Gilbert E, Grinter C, Gonzalez V, Horsley P, Lee S, Maier C, Miko I, Morris P, Oboyski P, Pierce N, Poelen J, Scott V, Smith M, Talamas E, Tsutsui N, Tucker E (2021) Announcing Big-Bee: An initiative to promote understanding of bees through image and trait digitization. Biodiversity Information Science and Standards 5 https://doi.org/10.3897/biss.5.74037
- Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone S, Scheuermann RH, Shah N, Whetzel PL, Lewis S (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25 (11): 1251-1255. https://doi.org/10.1038/nbt1346
- Stefancsik R, Balhoff J, Balk M, Ball R, Bello S, Caron A, Chesler E, de Souza V, Gehrke S, Haendel M, Harris L, Harris N, Ibrahim A, Koehler S, Matentzoglu N, McMurry J, Mungall C, Munoz-Torres M, Putman T, Robinson P, Smedley D, Sollis E, Thessen A, Vasilevsky N, Walton D, Osumi-Sutherland D (2023) The Ontology of Biological Attributes (OBA)—computational traits for the life sciences. Mammalian Genome 34 (3): 364-378. https://doi.org/10.1007/s00335-023-09992-1
- Tarasov S, Mikó I, Yoder MJ, Uyeda JC (2019) PARAMO: A Pipeline for Reconstructing Ancestral Anatomies Using Ontologies and Stochastic Mapping. Insect Systematics and Diversity 3 (6). https://doi.org/10.1093/isd/ixz009
- Thessen A, Walls R, Vogt L, Singer J, Warren R, Buttigieg PL, Balhoff J, Mungall C, McGuinness D, Stucky B, Yoder M, Haendel M (2020) Transforming the study of organisms: Phenomic data models and knowledge bases. PLOS Computational Biology 16 (11). https://doi.org/10.1371/journal.pcbi.1008376
- Thewlis J, Albanie S, Bilen H, Vedaldi A (2019) Unsupervised learning of landmarks by descriptor vector exchange. Proceedings of the IEEE/CVF International Conference on Computer Vision. https://doi.org/10.1109/ICCV.2019.00646
- Van Belleghem S, Papa R, Ortiz-Zuazaga H, Hendrickx F, Jiggins C, Owen McMillan W, Counterman B (2017) patternize: An R package for quantifying colour pattern variation. Methods in Ecology and Evolution 9 (2): 390-398. https://doi.org/10.1111/2041-210x.12853
- Violle C, Navas M, Vile D, Kazakou E, Fortunel C, Hummel I, Garnier E (2007) Let the concept of trait be functional! Oikos 116 (5): 882-892. https://doi.org/10.1111/j.0030-1299.2007.15559.x

- Vogt L (2019) Organizing phenotypic data—a semantic data model for anatomy. Journal of Biomedical Semantics 10 (1). https://doi.org/10.1186/s13326-019-0204-6
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7 (1). https://doi.org/10.1371/journal.pone.0029715
- Wilkinson M, Dumontier M, Aalbersberg I, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3: 160018. https://doi.org/10.1038/sdata.2016.18
- Yoder M, Mikó I, Seltmann K, Bertone M, Deans A (2010) A Gross Anatomy Ontology for Hymenoptera. PLoS ONE 5 (12). https://doi.org/10.1371/journal.pone.0015991

Endnotes

*1 https://phenoscape.org/
*2 https://github.com/phenoscape/TraitFest-2023/wiki/Participants
*3 https://github.com/phenoscape/TraitFest-2023
*4 https://github.com/phenoscape/TraitFest-2023/wiki/
*5 https://hackmd.io/ENiGYhDvT0a5ryjDdDOMEg
*6 https://github.com/phenoscape/TraitFest-2023/wiki/Code-of-Conduct
*7 https://github.com/phenoscape/TraitFest-2023/wiki/Agenda
*8 https://phenex.phenoscape.org/
*9 https://github.com/sergeitarasov/PhenoScript
*10 http://obofoundry.org/ontology/aism.html
*11 https://github.com/phenoscape/TraitFest-2023/wiki/Subgroups
*12 http://obofoundry.org/ontology/uberon.html
*13 https://obofoundry.org/ontology/hao.html
*14 https://github.com/phenoscape/pheno-repo
*15 www.gbif.org/species/5042859
*16 www.gbif.org/species/1339752
*17 www.gbif.org/species/4681203
*18 www.gbif.org/species/1334167
*19 www.w3.org/OWL/
*20 www.ontotext.com/knowledgehub/fundamentals/what-is-rdf/
*21 www.w3.org/TR/rdf-sparql-query/
*22 http://robot.obolibrary.org/template
*23 http://rs.tdwg.org/dwc/terms/catalogNumber
*24 http://tiny.cc/h3qavz
*25 https://www.hymenopterists.org/2023-congress/
*26 https://openai.com/blog/chatgpt
*27 https://github.com/megbalk/lit-bryo
*28 https://github.com/allenai/pdffigures2
*29 https://github.com/megbalk/bryo
*30 http://obofoundry.org/ontology/oba.html
*31 https://github.com/INCATools/ubergraph#sparql-endpoint
*32 https://github.com/phenoscape/TraitFest-2023/issues

*33 https://github.com/phenoscape/rphenoscape
*34 https://github.com/INCATools/ubergraph
*35 https://github.com/phenoscape/ubeRsim
*36 https://en.wikipedia.org/wiki/Jaccard_index#Tanimoto
*37 https://en.wikipedia.org/wiki/Cosine_similarity#Definition
*38 https://rdrr.io/cran/phytools/man/contMap.html
*39 https://github.com/tdwg/ac?tab=readme-ov-file#audiovisual-core-ac

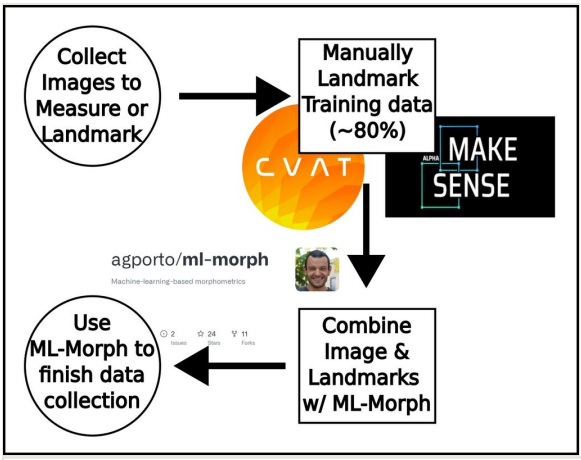

Figure 1.

Simplified workflow of landmark automation pipeline. Logos are included for Computer Vision Annotation Tool (CVAT) and Make Sense AI, free annotation tools that easily feed directly into the pipeline. This project uses the ML-morph tool (Porto and Voje 2020) and so we include a reference image to Porto's github repository for ML-morph.

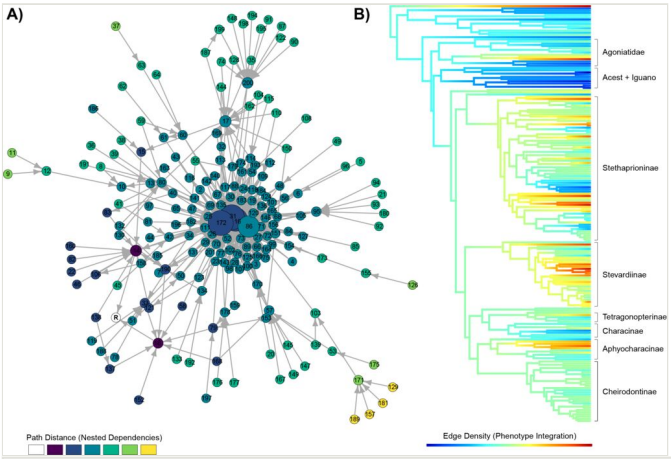

Figure 2.

A) Graph representation of anatomical entities and their dependencies, obtained from the Phenoscape KB, for fish taxa in the family Characidae. **B)** Ancestral state reconstruction showing the evolution of phenotype integration within Characidae. Integration values for each species (tip) were obtained by calculating the edge density of each species subgraph (i.e. subgraph including only anatomical entities present in a species).

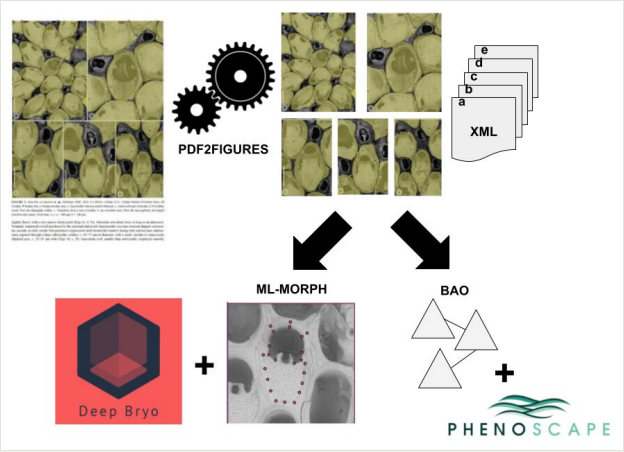

Figure 3.

Workflow of trait extraction from figures from literature. Figures from PDFs are extracted using pdf2figures. This results in images and xml files of their captions. We then extract trait terms and species names for the Byrozoa ontology, which then feeds into Phenoscape to build trait presence-absence matrices. The extracted images are fed into the machine-learning programmes DeepBryo and ML-morph to automatically annotate images while maintaining metadata from the figure caption.

Table 1.

Terms and identifiers for ten morphological structures of bryozoans in the new Bryozoan Attribute Ontology.

Term	**Identifier**
'pore chamber'	BRYO:0000001
'pore chamber cavity'	BRYO:0000002
'pore chamber plate'	BRYO:0000003
'primary orifice'	BRYO:0000004
'secondary orifice'	BRYO:0000005
'lophophore orifice'	BYRO:0000006
'lophophore opening'	BRYO:0000007
'lophophore feeding organ'	BRYO:0000008
'lophophorate feeding system'	BRYO:0000009
ovicell	BRYO:0000010