Research Ideas and Outcomes 10: e134840
doi: 10.3897/rio.10.e134840

Workshop Report

Community-driven enhancement of information ecosystems for the discovery and use of paleontological specimen data: Stakeholder engagement workshop

Talia Karim[‡], Erica Krimmel[§], Holly Little[|], Lindsay J. Walker[¶]

‡ University of Colorado Boulder, Boulder, United States of America
§ independent, Sacramento, United States of America
| Smithsonian National Museum of Natural History, Washington, DC, United States of America
¶ Arizona State University, Tempe, United States of America

Corresponding author: Erica Krimmel (ekrimmel@gmail.com)

Reviewable v 1

Received: 15 Aug 2024 | Published: 28 Aug 2024

Citation: Karim T, Krimmel E, Little H, Walker LJ (2024) Community-driven enhancement of information ecosystems for the discovery and use of paleontological specimen data: Stakeholder engagement workshop. Research Ideas and Outcomes 10: e134840. https://doi.org/10.3897/rio.10.e134840

Abstract

A stakeholder engagement workshop was held in May 2024 as part of the "Community-driven enhancement of information ecosystems for the discovery and use of paleontological specimen data" project, which is funded under the United States National Science Foundation (NSF) Geosciences Open Science Ecosystem (GEO OSE) program. This report describes the activites and outcomes of the workshop.

Keywords

paleontology, palaeontology, fossil, geology, biodiversity, collection, natural history collection, specimen

Date and place

This workshop took place from 14-15 May 2024 at the University of Colorado Boulder (CU) and was hosted by the CU Museum of Natural History.

List of participants

The twenty-four participants of this workshop (Table 1) were a mix of professionals focusing on research, collections care, and/or informatics in the paleontological domain (Fig. 1). Participants represented a range of career stages – including four graduate students – and a diversity of focal areas within the domain. Of the 12 participants who completed a demographics survey, 75% self-identified as female, 25% as Hispanic or Latino/Latina/Latinx, 25% with a racial identity other than White and 17% with a disability.

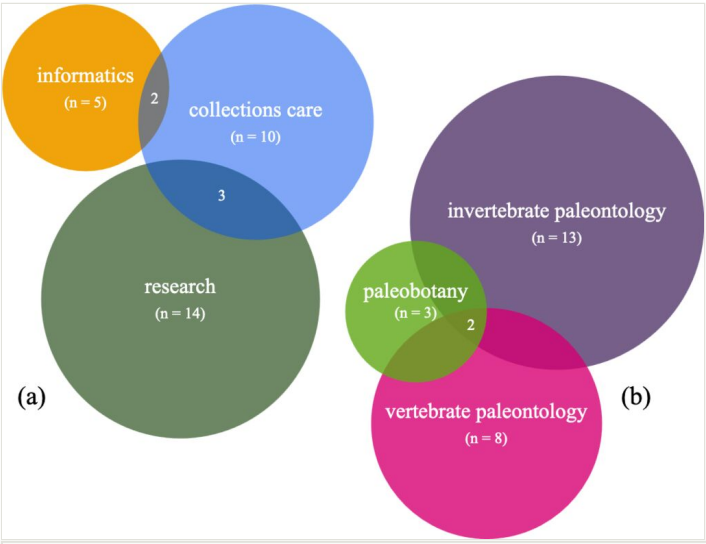

Figure 1. doi

Focal areas of workshop participants grouped by (a) professional training, and (b) taxonomic expertise.

Table 1.

List of on-site workshop participants. Workshop organizers are indicated with an asterisk (*).

Name	Institutional affiliation	Role or title
Alex Cano	Smithsonian National Museum of Natural History	Data Specialist
Amanda Millhouse	Smithsonian National Museum of Natural History	Deputy Collections Manager of Vertebrate Paleontology

Name	Institutional affiliation	Role or title
Carl Simpson	University of Colorado Boulder, Museum of Natural History	Assistant Professor of Geological Sciences | Curator of Invertebrate Paleontology
Casey Thater	University of Colorado Boulder, Museum of Natural History	Graduate Student
Chrissy Garcia	Stanford University	Geoscience Specimen Collection Curator and Manager
Corinne Myers	University of New Mexico	Associate Professor of Earth and Planetary Sciences
Ellen Currano	University of Wyoming	Professor of Paleobotany
Erica Krimmel*	independent	Information Scientist
Holly Little*	Smithsonian National Museum of Natural History	Informatics Manager of Paleobiology
Jacob Van Veldhuizen	University of Colorado Boulder, Museum of Natural History	Collections Manager of Vertebrate Paleontology
Jaelyn Eberle	University of Colorado Boulder, Museum of Natural History	Professor of Geological Sciences | Curator of Vertebrate Paleontology
Kit Lewers	University of Colorado Boulder	Graduate Student
Lindsay Walker*	Arizona State University	Symbiota Support Hub Community Manager
Melanie Hopkins	American Museum of Natural History	Curator of Invertebrate Paleontology
Nancy Stevens	University of Colorado Boulder, Museum of Natural History	Director | Professor of Anthropology
Natalia Lopez-Carranza	University of Kansas, Biodiversity Institute	Collection Manager – Invertebrate Paleontology
Nicole McGee	University of Colorado Boulder, Museum of Natural History	Graduate Student
Pat O'Connor	Ohio University	Professor of Anatomy and Neuroscience
Pedro Monarrez	University of California Los Angeles	Recruitment, Outreach, Diversity, Equity, and Inclusion Coordinator for the Department of Earth, Planetary, and Space Sciences
Sarah Leventhal	University of Colorado Boulder, Museum of Natural History	Graduate Student
Simon Goring	University of Wisconsin – Madison	Assistant Scientist
Stewart Edie	Smithsonian National Museum of Natural History	Research Geologist | Curator of Fossil Bivalvia

Name	Institutional affiliation	Role or title
Talia Karim*	University of Colorado Boulder, Museum of Natural History	Collections Manager of Invertebrate Paleontology and Paleobotany
Will Taylor	University of Colorado Boulder, Museum of Natural History	Assistant Professor of Anthropology | Curator of Archaeology

Introduction

This workshop is part of the "Community-driven enhancement of information ecosystems for the discovery and use of paleontological specimen data" project, which is funded under the United States National Science Foundation (NSF) Geosciences Open Science Ecosystem (GEO OSE) program. The goal of the project is to support transformational and translational research in the geo- and biosciences by driving development in the open data landscape, by improving discoverability and use of paleontological specimen data through community engagement and collaboration. Project personnel are actively coordinating with partners throughout the larger data ecosystem, including via two in-person workshops, of which this is the first.

At the intersection of geo- and bioscience, paleontology is an inherently interdisciplinary field and one with impactful research. The ever-growing climate crisis, for one example, highlights a need to understand how taxa reacted to changes in Earth's history, and underscores the importance of examining patterns from deep time into the modern. Over the last decade, the United States paleontology collections community has invested heavily in the digitization of primary specimen data, including over $10 million funded through the NSF Advancing Digitization of Biodiversity Collections (ADBC) program*[1]. These data are now accessible on open science platforms such as the Global Biodiversity Information Facility (GBIF)*[2] and iDigBio. However, GBIF and iDigBio were developed primarily for modern (neontological) biodiversity data. The resulting cyberinfrastructure gaps obscure critically useful primary data from paleontology collections, and inhibit integration between open science resources operating in geoscience vs. bioscience domains. This project is evaluating the existing technical landscape and laying the foundation for building out a network of FAIR, CARE, and research-ready data accessible via TRUSTed repositories (see, respectively, Wilkinson et al. (2016), Carroll et al. (2020), Lin et al. (2020)). This workshop provided an important opportunity to connect with stakeholders in the paleontological collections and research user communities.

Aims of the workshop

The desired outcomes of this stakeholder engagement workshop were:

- to collaboratively outline the needs of research and collections communities related to sharing and utilizing fossil collections data, and

- to discuss how these needs relate to existing and desired cyberinfrastructure.

By engaging a broad spectrum of individuals who interact with paleontological collections data in different ways, workshop organizers hoped to build a shared understanding of needs. As a component of the overarching project, these outcomes form the basis for advanced investigations into cyberinfrastructure needs and potential solutions that will be explored during the latter part of 2024 and into 2025.

Activities

After a welcome from Nancy Stevens, director of the CU Museum of Natural History Collections, the workshop kicked off with icebreaker activities designed to set an active and productive, yet informal, tone. Breakout group introduction discussions built trust and encouraged participants to learn about each other as individuals by posing questions like: What is something (non-work) that you have accomplished recently and are proud of? What keeps you up at night, good or bad (work-wise)? What are your hopes and dreams for paleo data?

Data Use Spotlights

Throughout the two days, most workshop participants shared briefly about their work via an activity we called a "Data Use Spotlight." Instructions for this activity were to prepare one slide to illustrate how they use fossil data for their work (Fig. 2). These helped set the context, as well as provided more insight for all participants into the presenter's area of expertise. The Data Use Spotlights were a valuable part of the workshop in unanticipated ways too. For instance, they repeatedly opened the door to fruitful discussions, either directly as part of the workshop agenda, or indirectly via conversations during breaks and beyond. In several cases, these presentations provided the basis for questions used in other workshop activities, and inspired potential collaborations amongst participants.

Resources round-up for the paleo data ecosystem map

Workshop organizers presented an overview of their vision for a paleo data ecosystem map, contextualized as the universe of resources we use to do our work and how these resources interact with each other. Creating this map will involve modeling the existing information and systems landscape by characterizing various resources (concepts, systems, platforms, mechanisms, drivers, tools, documentation, standards, etc.), and specifically addressing their use for fossil data. The resulting map will be a tool with entry points for multiple audiences, including new members to the community, members working in specific sectors, and members working to integrate initiatives and systems.

With the context provided by this overview, participants worked collaboratively to list resources they use on giant sticky notes. Essential resources were then flagged with pink sticky notes, and additional sticky note colors were added to capture comments about how participants use the resource, and how it might be tagged in the envisioned

ecosystem map (Fig. 3). Workshop participant responses subsequently were tallied (Fig. 4) and documented for future integration into the project's ecosystem map. This activity provided concrete information on what resources (analog and digital) researchers are aware of and their general frequency of use.

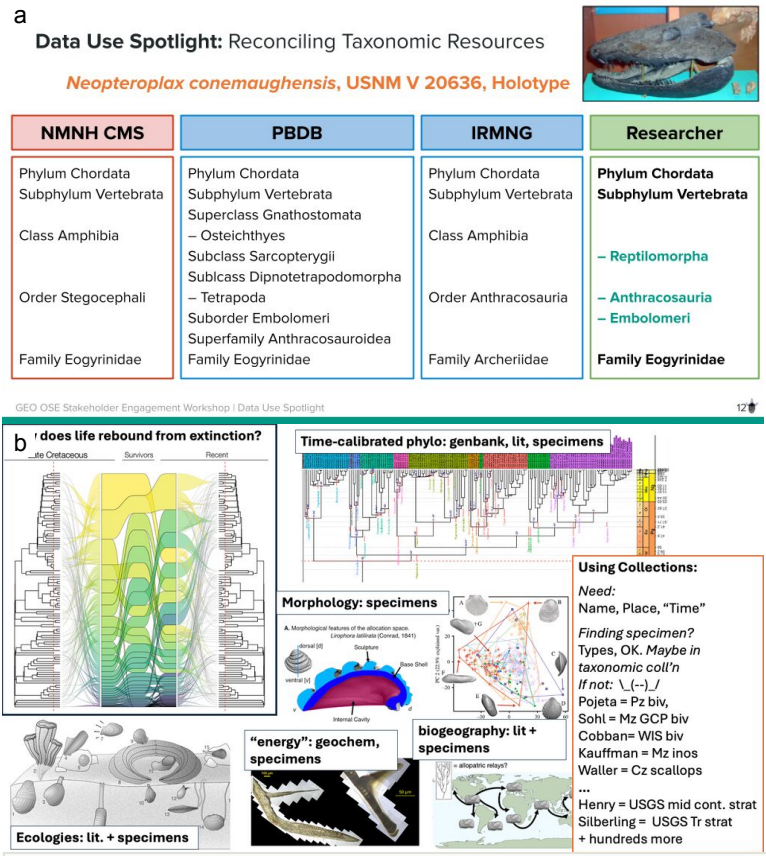

Figure 2.

Example "Data Use Spotlight" slides.

a: Amanda Millhouse spoke about reconciling taxonomic resources in the context of collections management (CC BY 4.0, Amanda Millhouse). doi

b: Stewart Edie spoke about his research on species diversity after bottleneck events and the challenges of discovering valuable data from collections (CC BY 4.0, Stewart Edie). doi

Connecting specimens into the paleo data ecosystem

This activity began with providing an overview of paleo collections as core research infrastructure (NASEM 2020), highlighting that the distributed nature of collections is both advantageous and challenging. Digitization over the past couple of decades has allowed us to build a vision of paleo collections as being distributed yet connected; one of the goals of this project is to identify how we might continue to improve this vision in practice.

Figure 3. doi

"Resources round-up" sticky notes visible on the far wall of the workshop room. Note that pink stickies mark essential resources.

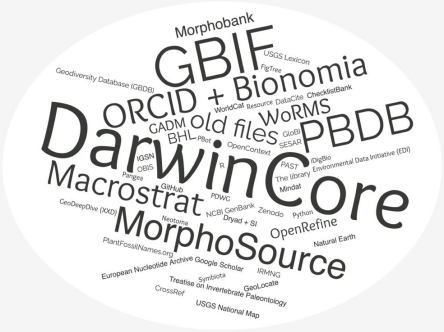

Figure 4. doi

Word cloud of all resources listed in the "Resources round-up," visualized such that size is proportional to value, i.e. the number of pink sticky notes added to a given resource.

Considering both physical (in person or via loan) and digital (data and/or media) access, this project is building a conceptual data model for paleo specimens where we can classify different types of data and describe the attributes of and relationships between classes. We expect that this model may be useful for tasks such as:

- Determining where different classes of data come from
- Considering who uses different classes of data, and for what
- Asking who curates and maintains different classes of data

- Discovering potential roadblocks to data capture, mobilization, and/or use
- Evaluating the relationships between data classes
- Questioning whether we are capturing and sharing the right attributes for each class of data
- Identifying where our data standards need to evolve

In small groups, workshop participants sketched out connections to the data they might need in order to use a given specimen for research (Fig. 5). Considering these sketches, participants discussed as a group: What data are tricky to find? What data are tricky to use? What data are most critical for your work? Where do you spend the most time? This activity allowed workshop organizers to ground truth prior conceptions about how researchers perceive data related to fossil specimens, and to identify information that is either missing or inaccessible.

Mapping data pipelines from source to science

The final workshop activity focused on developing a better understanding of the research data pipeline. Workshop organizers asked participants to write down research questions (old, new, previously examined, or unsolved) on sticky notes. These were grouped thematically into three large clusters and one question from each group was chosen as an exemplar research question to explore. Participants were asked to map out all the steps they would do in order to answer their exemplar question and where there could be resource gaps or challenges that would inhibit the research process. This allowed the group to better understand how fossils and associated data are utilized and accessed as part of the research data pipeline.

Group A focused on "How do we find new fossils?" and identified the key data points needed to answer various iterations of this question (Fig. 6a, b). They then discussed existing resources that can be used to discover those data points and the challenges in being able to access or utilize the data fully. For example, some existing resources (e.g. MacroStrat) do not provide the full scope or ideal format for the data needed, while others (e.g. Geobiodiversity Database) are inconsistently accessible. Ultimately, this group noted that infrastructure and practices that better support linked data would be the ideal solution for enabling the research data pipline necessary for addressing questions in this theme.

Group B focused on biogeography, comparing niche dimensions with phylogeny (Fig. 6c). They identified the need to integrate a phylogeny with geographic occurrence and environmental data, but also noted that there are trust issues with occurrence data and a high cost to obtaining environmental data. The group posited that occurrence data at the species inventory level (versus the individual organism level) might be more practically useful for answering biogeographical questions.

Group C focused on trait data, specifically, looking at trait selectivity to predict extinction risk across a geologic time boundary, for example, the Cretaceous-Paleogene Boundary (Fig. 6d). Initially, they wanted to know if trait data already exist somewhere by examining

literature and collections. This group mapped a research pipeline relying more on physical collections and interpersonal information exchange to discover data, and less on publicly available digital resources, although MorphoSource and MorphoBank were considered as possible data sources. Some discussion of data cleaning and consolidation revealed different understanding of these tasks from the collections management versus research perspectives. They recognized the importance of making trait data FAIR at the point of capture, as well as the need for standards to share these types of specialized research data.

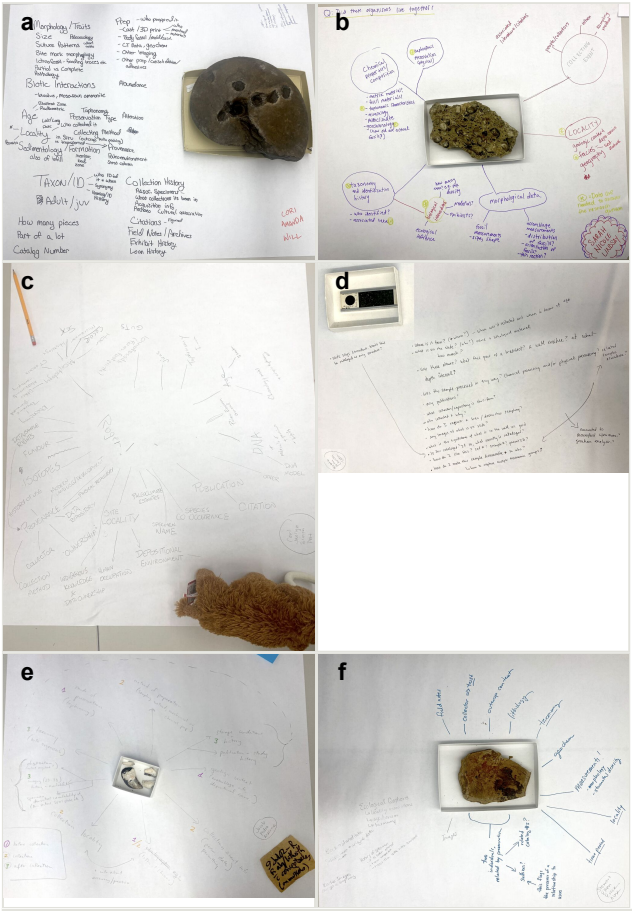

Figure 5.

Results from the activity, "Connecting specimens into the paleo data ecosystem," where each group worked with a different specimen.

a: Ammonite specimen doi
b: Invertebrate slab specimen doi
c: Mammoth specimen doi
d: Microfossil specimen doi
e: Mollusk specimen doi
f: Plant specimen doi

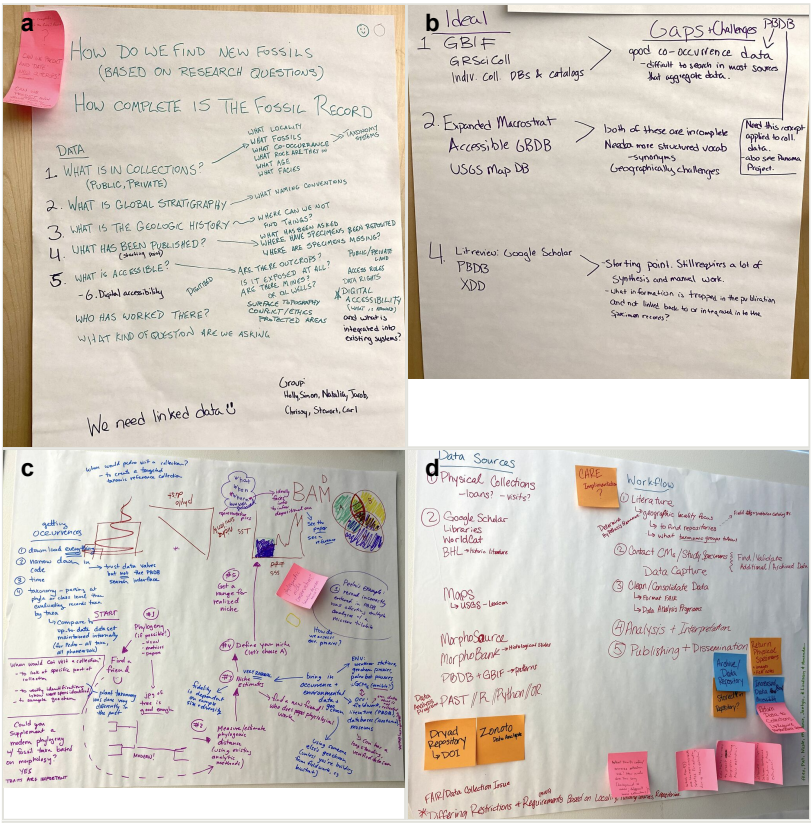

Figure 6.

Results from the activity, "Mapping data pipelines from source to science," where each group evaluated a different research question.

a: Group A (part I) doi
b: Group A (part II) doi
c: Group B doi
d: Group C doi

As with previous activities, this one provided workshop organizers with invaluable perspective about how researchers perceive and use fossil specimen data, both digital and analog. The diagrams resulting from this activity will inform future work on the overarching project.

Participant feedback

All participants were asked to provide anonymous feedback on the workshop via a brief survey, which was separate from a demographics survey. Eleven people responded, representing slightly over half of the 20 workshop participants (workshop organizers did not participate in this survey). Feedback provided in the survey was overwhelmingly positive (Fig. 7). In a free-response question, participants noted that they appreciated the

opportunity this workshop provided to connect with others and gain a better understanding of the challenges their colleagues face either mobilizing or using paleo data. They also found discussions about specific resources and initiatives to be valuable. Survey respondents wanted to know more about both tractable topics (e.g. funding opportunities, specific resources mentioned) and complex topics (e.g. how to better work together across research and collections care, how to link data in an ideal world). Participants uninimously agreed that this workshop met their expectations, and additionally provided constructive criticism about the various workshop activities (Fig. 8).

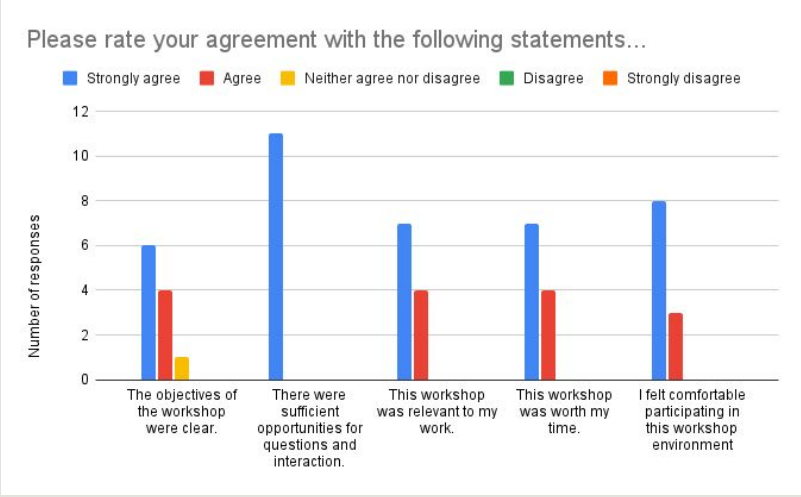

Figure 7. doi

Categorical responses from the post-workshop anonymous feedback survey.

Figure 8. doi

Responses from the post-workshop anonymous feedback survey for the question, "What workshop activities do you think either worked well or could be improved?"

Key outcomes and discussions

Throughout the workshop, participants highlighted critical themes that align with the big-picture objectives of this project.

Fitness-for-use of specimen-based data available on aggregators (e.g. GBIF, iDigBio) was one such recurrent theme. Discussions touched on use of specimen images for diverse research purposes, digitization of data "on demand," the necessity of species-level taxonomic identifications, and duplication of occurrence records. Participants were particularly interested in considering the “why” of collecting, as knowing why something was collected could inform its fitness-for-use in other applications.

Fitness-for-use ties directly into another theme, *data availability*. Workshop participants had many discussions focused on what data are available, what data are not, and (if not) why not? On a specimen level, participants discussed availability of digitized trait data, which are typically not stored with the specimen record itself, or shared on data aggregators for fossil specimens. On a broader level, participants explored the idea of sharing minimal data to improve discoverability of larger collections where specimen-level digitization is an unreasonable target (e.g. an institution might share inventory data via the Latimer Core standard to let researchers know about all brachiopods collected by a particular person). Such minimal data might be the entry point for digitization "on demand" of data at the specimen-level. On a human level, participants discussed how much institutional knowledge is held by collection staff, and how best to capture that before individuals retire or move on.

Finally, the *capacity* to make collections data fit-for-use and available came up constantly. Several participants shared that they were the only people at their institutions managing those collections. For others, the scope of digitizing legacy data in their collections is so vast that multiple additional trained staff would be needed to address the issue. Everyone was concerned about how we might try to future-proof existing research datasets and databases. Who is going to maintain these key resources in the future when we barely have the capacity and funding to do it now?

All three of these themes emphasize that *humans* are at the center of research and collections. In planning this workshop, we attempted to be people-centric. Built-in flexibility in the agenda allowed participants to have time for discussions when a topic emerged that sparked group interest. Similarly, providing longer lunch and coffee breaks facilitated unstructured discussion and allowed people to think, chat, and explore ideas organically. Concrete results from the workshop activities are valuable to the overarching project, but equally so was laying the groundwork for continuing to have productive and collaborative conversations with the group of people who participated. Building a shared understanding of the needs of research and collections communities related to fossil data is an ongoing process, and one that is essential to envisioning solutions.

To conclude, this stakeholder engagement workshop brought together a group of professionals with varied skillsets, perspectives, and end-use goals for digitized fossil

collections data. In two days, the group provided critical feedback to defining the essential elements of the vast landscape (or ecosystem) of research resources available to the paleontological community, modeled data pipelines based on real-life questions in paleontological research, and became better acquainted with the data needs, uses, and workflows of colleagues working in other sectors of the paleontological domain. While much progress remains to be accomplished, the outcomes of this workshop underscore the need for the paleontological research, collections, and informatics specialists to collaboratively define solutions for data pipelines through people-centric initiatives.

Acknowledgements

Thank you sincerely to all those who participated in this workshop and made it the success that it was! Extra thanks to Kit Lewers, Alex Cano, Nicole McGee, Jerah Brewster, and Sam Eads for their help with lunchtime logistics.

Funding program

This workshop is part of the project, “Community-driven enhancement of information ecosystems for the discovery and use of paleontological specimen data,” which is funded by the NSF GEO OSE program under the following awards: 2324688, 2324689, 2324690.

Hosting institution

University of Colorado Boulder

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Carroll SR, Garba I, Figueroa-Rodríguez O, Holbrook J, Lovett R, Materechera S, Parsons M, Raseroka K, Rodriguez-Lonebear D, Rowe R, Sara R, Walker J, Anderson J, Hudson M, et al. (2020) The CARE Principles for Indigenous Data Governance. Data Science Journal 19 https://doi.org/10.5334/dsj-2020-043
- Lin D, Crabtree J, Dillo I, Downs R, Edmunds R, Giaretta D, De Giusti M, L’Hours H, Hugo W, Jenkyns R, Khodiyar V, Martone M, Mokrane M, Navale V, Petters J, Sierman B, Sokolova D, Stockhause M, Westbrook J, et al. (2020) The TRUST Principles for digital repositories. Scientific Data 7 (1). https://doi.org/10.1038/s41597-020-0486-7
- NASEM (2020) Biological Collections: Ensuring Critical Research and Education for the 21st Century. The National Academies Press.https://doi.org/10.17226/25592

- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18

Endnotes

*1 NSF-funded "Thematic Collections Networks" including fossil specimens: Fossil Insect Collaborative (2013-2020), PaleoNICHES (2012-2015), EPICC (2015-2020), Cretaceous World (2016-2023), Pteridological Collections Consortium (2018-2023). Monetary amount acquired from NSF's Award Search.

*2 As of this writing (2024-06-04), there are 8,917,071 occurrence records in the GBIF data portal where *basisOfRecord* = "FossilSpecimen".