Developing a vocabulary and ontology for modeling insect natural history data: example data, use cases, and competency questions

Brian J. Stucky[‡], James P. Balhoff[§], Narayani Barve[‡], Vijay Barve[‡], Laura Brenskelle[‡], Matthew H. Brush[|], Gregory A Dahlem[¶], James D. J. Gilbert[#], Akito Y. Kawahara[‡,¤], Oliver Keller[¤], Andrea Lucky[¤], Peter J. Mayhew[«], David Plotkin[‡], Katja C. Seltmann[»], Elijah Talamas[^], Gaurav Vaidya[‡], Ramona Walls[˅], Matt Yoder[¦], Guanyang Zhang[‡], Rob Guralnick[‡]

‡ Florida Museum of Natural History, University of Florida, Gainesville, FL, United States of America
§ Renaissance Computing Institute, University of North Carolina, Chapel Hill, NC, United States of America
| Oregon Health and Science University, Portland, OR, United States of America
¶ Department of Biological Sciences, Northern Kentucky University, Highland Heights, KY, United States of America
# Department of Biological and Marine Sciences, University of Hull, Hull, United Kingdom
¤ Entomology and Nematology Department, University of Florida, Gainesville, FL, United States of America
« Department of Biology, University of York, York, United Kingdom
» Cheadle Center for Biodiversity and Ecological Restoration, University of California, Santa Barbara, Santa Barbara, CA, United States of America
^ Florida Department of Agriculture and Consumer Services, Gainesville, FL, United States of America
˅ Bio5 and CyVerse, University of Arizona, Tucson, AZ, United States of America
¦ Species File Group, Illinois Natural History Survey, University of Illinois, Champaign, IL, United States of America

Corresponding author: Brian J. Stucky (stucky.brian@gmail.com)

Academic editor: Vincent Smith

Abstract

Insects are possibly the most taxonomically and ecologically diverse class of multicellular organisms on Earth. Consequently, they provide nearly unlimited opportunities to develop and test ecological and evolutionary hypotheses. Currently, however, large-scale studies of insect ecology, behavior, and trait evolution are impeded by the difficulty in obtaining and analyzing data derived from natural history observations of insects. These data are typically highly heterogeneous and widely scattered among many sources, which makes developing robust information systems to aggregate and disseminate them a significant challenge. As a step towards this goal, we report initial results of a new effort to develop a standardized vocabulary and ontology for insect natural history data. In particular, we describe a new database of representative insect natural history data derived from multiple sources (but focused on data from specimens in biological collections), an analysis of the abstract conceptual areas required for a comprehensive ontology of insect natural history data, and a database of use cases and competency questions to guide the development of data systems for insect natural history data. We also discuss data modeling and technology-related challenges that must be overcome to implement robust integration of insect natural history data.

©

Keywords

insects, natural history, biodiversity informatics, ontology, data modeling

Introduction

Insects are possibly the most diverse class of multicellular organisms on Earth, not only in sheer number of species, but also in terms of ecological diversity (Grimaldi and Engel 2005, Larsen et al. 2017). For example, insects encompass just about every sort of trophic strategy known in animals, including herbivory, scavenging, predation, parasitism, and parasitoidism. In some cases, all of these strategies are found within a single taxonomic family (e.g., Disney 1994, Marshall 2012, Rainford and Mayhew 2015). Thus, insects present boundless opportunities to test hypotheses about the ecology and evolution of feeding behaviors, species interactions, habitat associations, and much more. Actually realizing this potential within a scientific study, however, is quite challenging because of the difficulty in obtaining, integrating, and analyzing suitable natural history data.

Currently, data about the natural history of insects are widely scattered among a multitude of sources, including labels on specimens in biological collections, specialized (and often obscure) publications, field notebooks, and taxon-specific databases. Thus, finding relevant natural history data for a given insect species can be a daunting task. Furthermore, insect natural history data are highly heterogeneous. For example, they commonly differ in observational methodology (e.g., observations in the field versus in the lab), observational detail (e.g., differences in temporal resolution or certainty of biotic associations), or in the terminology used by the observers. Aggregating these data so that they can be analyzed and disseminated efficiently, without information loss, is a major informatics challenge.

A critical step towards meeting this challenge is developing comprehensive standards to guide the design and implementation of data systems for aggregating insect natural history data. To support robust data integration, these standards need to include two major components: first, a well-defined vocabulary of natural history terms that is suitable for recording natural history observations across all insect taxa and, second, an ontology that provides computable semantics for the vocabulary so that computers can understand how the terms in the vocabulary relate to one another (ontologies are described in the next section). Such data standards can have a major impact on large-scale biodiversity science, as exemplified by the success of the "Darwin Core" vocabulary for aggregating and exchanging species occurrence data (Wieczorek et al. 2012).

Here, we report initial results of a new effort to develop a standardized vocabulary and ontology for insect natural history data, an effort that was initiated at a three-day workshop, held at the University of Florida from 1 May to 1 June 2018, that convened

entomologists, computer scientists, and data modelers. Although work on a draft ontology is still in progress, in this short communication we describe several key results of our work so far that are likely to be of broader interest, including an analysis of high-level ontology concept areas, a conceptually comprehensive database of example insect natural history data, and a database of ontology use cases and ontology competency questions.

To make this work tractable, we have mostly focused on natural history information from specimens in collections, with taxonomic scope limited to the five mega-diverse insect orders (Hemiptera, Coleoptera, Diptera, Lepidoptera, Hymenoptera), which include the vast majority of insect species and ecological diversity (Grimaldi and Engel 2005). We also excluded natural history information inferred from fossil material. There is considerable overlap in content between insect natural history data from specimen labels and from other sources (e.g., literature), so much of our work will be easily adaptable to information about other insect orders or information from sources other than specimen labels. Looking even further ahead, we anticipate that an ontology for insect natural history data could eventually serve as a foundation for developing a broader ontology for natural history data that also includes other groups of animals.

Before turning to discussion of our vocabulary and ontology development efforts, we recognize that many readers might have little experience with ontologies, so we briefly introduce ontologies and why they are important for integrating natural history data.

A (very) brief introduction to ontologies

An *ontology*, as the term is used in computer and information science, is an explicit, precise, machine-interpretable conceptualization of some knowledge domain. Although we do not have space in this manuscript to provide a detailed introduction to ontologies, we will try to provide some intuition by way of a simple example. Suppose we have two natural history observations: observation 1 asserts that an individual of species *A* was a parasitoid of an individual of species *B* and observation 2 asserts that an individual of species *C* was a predator of species *B* (Fig. 1). Now, suppose we have a database that includes these two observations (and potentially many more), and we wish to query the database to find all of the species that are known to use species *B* as a food source.

Given observations 1 and 2, a human biologist can easily infer that species *A* and *C* are both known to feed on species *B*, but a computer does not automatically understand that "parasitoid of" and "predator of" both imply trophic relationships. With an ontology, we can provide formal logic statements, called *axioms*, that allow a computer to make this inference. To continue with the example, we could write axioms that assert that the relationships "parasitoid of" and "predator of" are both special cases of a more general relationship called "feeds on" Fig. 1. Armed with this information, a computer could directly answer our question about which species use species *B* as a food source.

With only two observations and a few vocabulary terms, this might seem like a trivial accomplishment, but when we have hundreds, thousands, or even millions of heterogeneous natural history observations, with hundreds of logical relationships among the terms in a large vocabulary, ontologies make it possible to automate complex data integration and querying tasks that would be practically impossible for a human. Thus, ontologies are critical to any effort to develop robust systems for aggregating insect natural history data. Furthermore, although this brief discussion has focused on the value of ontologies for data aggregators and users, ontologies are also beneficial for data creators and providers because they provide a standardized vocabulary that, once adopted, makes an individual's or organization's data immediately interoperable with similar data from other sources. This, in turn, makes the data more likely to be used (and cited) by other researchers. For readers who wish to learn more about data modeling with ontologies, Allemang and Hendler (2011) provide a good introduction.

Development tasks, methods, and outcomes

We now return to discussion of the ontology design and development work initiated at the workshop, which has been organized around four major tasks: 1) assembly of example data; 2) analysis of example data and ontology scoping; 3) high-level ontology design and concept identification; and 4) identifying use cases (and users) and authoring ontology competency questions. We briefly describe each of these tasks and present the results of our work so far.

Assembly of example data

Insect natural history is an extremely broad domain, which means that identifying an appropriate scope for a new data vocabulary and ontology is not a simple task. Our approach to this problem was to assemble example natural history data, drawn from real data sources, for each of the five major insect orders. This served two purposes. First, examining a well-drawn set of example data is a practical method for delimiting the scope of a new vocabulary and ontology, and second, a good example dataset also provides valuable test cases for use during vocabulary and ontology development.

To generate the example dataset, we worked in five small groups. Each group was assigned one of of the five major insect orders, and we ensured that each group included at least one entomologist with expertise in the assigned order. Then, each group gathered example natural history data for their insect order, with the goal of compiling a concise dataset that represented the various kinds of natural history information recorded on specimen labels for each major insect order. We attempted to capture both the breadth of biological information and the range of observational detail found in label data. Although we focused on information from insect specimen labels, we also included some data from literature sources and online databases such as iNaturalist (https://www.inaturalist.org) and GloBI (Poelen et al. 2014). For data from specimen labels, we used specimens and labels with digital images available on iDigBio (Page et al. 2015)

whenever possible. Example data we gathered at the workshop were supplemented by additional example data that a few participants gathered both prior to and after the workshop.

Our final dataset includes 189 natural history observations covering a wide range of concepts and observation types (see next section). We expect that this dataset will have value to other researchers as well, so we have included it with this manuscript as two supplemental files, with one file formatted as a PDF document (Suppl. material 1) and one file in tabular comma-separated values (CSV) format (Suppl. material 2). Both of these files are also available in a public git repository hosted on GitLab which provides the example data in other formats, too, including styled HTML, Markdown, and a SQLite database (https://gitlab.com/stuckyb/inhd_ontology/tree/master/example_data).

Analysis of example data and ontology scoping

After assembling the example data, we used them to delimit the high-level scope of the new vocabulary and ontology. Again working in small groups, we analyzed the kinds of information contained in the example data, with each group focusing on one of the five major insect orders. For each order, we summarized the kinds of biological information that were observed (e.g., various multi-organism interactions, developmental data) and the ways in which the information was recorded (e.g., qualitative or quantitative). Then, we reconvened as a large group, each small group reported their findings, and we synthesized the results to arrive at a set of 10 high-level conceptual areas required for the final ontology (Table 1).

Together, these conceptual areas cover virtually all of the kinds of information contained in the example data we assembled, and we therefore propose that an ontology that provides suitable coverage of all 10 of these areas will be sufficient for modeling nearly all insect natural history data from specimen labels as well as a substantial proportion of insect natural history data from other sources, including literature-based data. This conclusion is dependent, of course, on the extent to which our example data capture the conceptual breadth and depth of all available insect natural history information. Although we were not able to formally evaluate this, given the collective entomological expertise of the workshop participants (many of whom have years of experience examining specimens and labels from entomology collections around the world) and the effort spent compiling example data, we are confident that we at least came close to achieving this goal for natural history data from insect specimen labels.

We also note that several of these conceptual areas overlap with the domains of extant ontologies, and in Table 2, we list the ontologies that are most relevant to each conceptual area. To ensure broad compatibility, reusability, and extensibility, we plan to use existing ontological resources wherever possible and contribute (or suggest) new entities for extant ontologies, when appropriate.

High-level ontology design and concept identification

Of the 10 conceptual areas we identified (Table 1), we determined that observations and observing processes, relationships and interactions, and positional (spatial) information were the most critical for developing an immediately useful vocabulary and ontology. Our decision to prioritize these areas was based on three considerations. First, observations and observing processes underlie *all* insect natural history data and encompass the crucial "who", "when", and "where" information about such data. Second, relationships and interactions are of broad scientific interest because they provide the raw ecological information needed for a wide variety of research topics (e.g., understanding trophic relationships, discovering potential disease vectors, or predicting the consequences of ecosystem changes). Third, we found that positional or spatial information is often included on specimen labels and in literature-based natural history observations, and we therefore concluded that even a minimal data standard should be able to capture such information. After prioritizing these three conceptual areas, we again worked in groups to begin sketching out data models (Simsion 2007, Simsion and Witt 2005) and ontology design patterns (Gangemi 2005) for all three areas and to identify the entities (concepts) to include in each conceptual area.

This initial design work revealed several critical data modeling challenges, the thorniest of which is the problem of recording metadata about natural history observations that include interactions between organisms. Such observations are common in natural history
data and include, for example, observations about feeding relationships, parasite/host relationships, courtship, and many more. As with any other natural history observation, it is important to be able to record metadata about interaction observations, such as who made the observations, when they occurred, and so on. Without plunging into too much technical detail, the central problem is that the technology most often used for implementing ontology-enabled data, the Resource Description Framework (RDF, Miller 2005), currently has poor support for expressing metadata about interactions or relationships (Hartig 2017). A number of workarounds have been proposed (e.g., Hartig 2017, Nguyen et al. 2014, Hernández et al. 2015), but most of them have undesirable consequences, such as artificially increasing database size, complicating query statements, or slowing query response times (Hartig 2017, Hernández et al. 2015). Our work on this is ongoing, and we are actively investigating several different implementation strategies.

A second important data modeling problem is the challenge of accurately capturing information about *what* organisms were observed, which means dealing with the myriad difficulties posed by the use of taxonomic names (Zermoglio et al. 2016, Hardisty et al. 2013, Remsen 2016, Pyle 2016, Patterson et al. 2016). These issues are especially severe when dealing with data about insects, simply because insects are so extraordinarily diverse: many species remain undescribed and specimens in collections are often not identified to species (indeed, for some diverse insect families, the *majority* of

specimens in a collection might not be identified to species). Relatively frequent – and sometimes dramatic – taxonomic changes mean that the names used in publications and labels can quickly become inaccurate or obsolete. These issues are certainly not unique to insect natural history data, and we have not attempted to add to the substantial work already done in this area (e.g., Franz et al. 2017, Franz and Peet 2009, Hardisty et al. 2013, Pyle 2016). For now, though, taxonomic integration remains a major challenge for virtually all biodiversity-related data aggregation efforts, and insect natural history data are no exception.

Identifying use cases and authoring ontology competency questions

The last major task of our preliminary design and development work was drafting detailed ontology competency questions and identifying potential users and user cases. Ontology competency questions (OCQs, Grüninger and Fox 1995, Ren et al. 2014) provide a means for testing an ontology by providing specific queries that an ontology (along with an associated database) ought to be able to answer. In other words, OCQs specify how an ontology will be used to ask questions of real data. Thus, writing OCQs goes hand-in-hand with determining an ontology's users and use cases. To give a couple of examples, OCQs for an ontology of insect natural history data might include, “On what substrates does species *A* lay its eggs?” or “Has species *B* been collected at artificial lights?”

To identify use cases and develop OCQs, we divided into three groups on the last day of the workshop, with each group working independently and recording their results. After the workshop, one of us (BJS) synthesied the results of each group’s efforts into a single, comprehensive set of use cases and OCQs. The use cases we identified cover seven main user groups or domains:

1. Entomology (e.g., insect collecting and rearing, forensic entomology).
2. Taxonomy and systematics (e.g., field guides, systematic revisions).
3. Ecology and evolutionary biology (e.g., disease ecology, comparative studies).
4. Conservation biology and natural resource management (e.g., ecological restoration, environmental monitoring).
5. Agriculture and forestry (e.g., identifying potential pest insects, identifying potential disease vectors).
6. Education (e.g., classroom education, public outreach).
7. The general public (e.g., researching garden pests and control agents, hobby insect collecting).

The full sets of use cases and OCQs are too large to report in the main text, so we instead provide them in Suppl. material 3. The use cases and OCQs are also available in a public git repository on GitLab, which includes a SQLite database of use cases and OCQs along with example queries (https://gitlab.com/stuckyb/inhd_ontology/tree/master/OCQs).

Conclusions

With the work and results reported in this paper, we have laid a foundation for ongoing efforts to design, develop, and implement a robust vocabulary and ontology for modeling insect natural history data. Our next immediate goals are to identify the best solution for dealing with the problem of interactions metadata, discussed above, and to produce and release a draft ontology implementation for public review. We welcome additional participants in these efforts; readers who would like to be involved should contact the corresponding author (BJS). In the meantime, we hope that the foundational work reported in this paper, including the comprehensive example dataset and OCQs, will prove useful to other researchers interested in the informatics challenges surrounding insect natural history data.

Acknowledgements

This work was supported by the National Science Foundation Postdoctoral Research Fellowship in Biology under grant no. 1612335 to BJS, a University of Florida Informatics Institute fellowship to BJS, and by an iDigBio workshop grant. We thank C. Bester, K. Love, and other iDigBio personnel who helped coordinate workshop logistics. We also thank four reviewers for their helpful and insightful comments.

Conflicts of interest

References

- Allemang D, Hendler JA (2011) Semantic Web for the Working Ontologist: Effective Modeling in Rdfs and Owl. 2nd Edition. Morgan Kaufmann/Elsevier, Waltham, MA, USA, 354 pp. [ISBN 978-0-12-385965-5] https://doi.org/10.1016/B978-0-12-385965-5.10016-0
- Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry JM, Davis A, Dolinski K, Dwight S, Eppig J, Harris M, Hill D, Issel-Tarver L, Kasarskis A, Lewis S, Matese J, Richardson J, Ringwald M, Rubin G, Sherlock G (2000) Gene Ontology: tool for the unification of biology. Nature Genetics 25 (1): 25-29. https://doi.org/10.1038/75556
- Ashburner M, Schriml L (2013) GAZ: An open source gazetteer constructed on ontological principles. URL: http://environmentontology.github.io/gaz/
- Buttigieg P, Morrison N, Smith B, Mungall CJ, Lewis SE, Consortium tE (2013) The environment ontology: contextualising biological and biomedical entities. Journal of Biomedical Semantics 4 (1): 43. https://doi.org/10.1186/2041-1480-4-43
- Buttigieg PL, Pafilis E, Lewis S, Schildhauer M, Walls R, Mungall C (2016) The environment ontology in 2016: Bridging domains with increased scope, semantic density, and interoperation. Journal of Biomedical Semantics 7 (1). https://doi.org/10.1186/s13326-016-0097-6

- Dahdul WM, Cui H, Mabee PM, Mungall CJ, Osumi-Sutherland D, Walls RL, Haendel MA (2014) Nose to tail, roots to shoots: spatial descriptors for phenotypic diversity in the Biological Spatial Ontology. Journal of Biomedical Semantics 5 (1): 34. https://doi.org/10.1186/2041-1480-5-34
- Disney RHL (1994) Scuttle Flies: The Phoridae. Springer, Dordrecht, Netherlands, 467 pp. https://doi.org/10.1007/978-94-011-1288-8
- Franz N, Zhang C, Lee J (2017) A logic approach to modelling nomenclatural change. Cladistics 34 (3): 336-357. https://doi.org/10.1111/cla.12201
- Franz NM, Peet RK (2009) Perspectives: Towards a language for mapping relationships among taxonomic concepts. Systematics and Biodiversity 7 (1): 5-20. https://doi.org/10.1017/s147720000800282x
- Gangemi A (2005) Ontology design patterns for semantic web content. The Semantic Web – ISWC 2005. Lecture Notes in Computer Science. Vol. 3729. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11574620_21
- Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research 32 (90001): 258D-261. https://doi.org/10.1093/nar/gkh036
- Gkoutos G, Schofield P, Hoehndorf R (2012) The Neurobehavior Ontology. In: Chesler E, Haendel M (Eds) International Review of Neurobiology: Bioinformatics of Behavior: Part 1. https://doi.org/10.1016/b978-0-12-388408-4.00004-6
- Grimaldi DA, Engel MS (2005) Evolution of the Insects. Cambridge University Press, Cambridge & New York, 755 pp. [ISBN 978-0-521-82149-0] https://doi.org/10.1163/187631205794761021
- Grüninger M, Fox MS (1995) The role of competency questions in enterprise engineering. In: Rolstadås A (Ed.) Benchmarking — Theory and Practice. IFIP Advances in Information and Communication Technology. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-34847-6_3
- Hardisty A, Roberts D, community Tbi (2013) A decadal view of biodiversity informatics: challenges and priorities. BMC Ecology 13 (1): 16. https://doi.org/10.1186/1472-6785-13-16
- Hartig O (2017) Foundations of RDF and SPARQL: An alternative approach to statement-level metadata in RDF. CEUR Workshop Proceedings 1912 URL: http://ceur-ws.org/Vol-1912/paper12.pdf
- Hernández D, Hogan A, Krötzsch M (2015) Reifying rdf: what works well with Wikidata? CEUR Workshop Proceedings 1457: 32-47. https://doi.org/10.1145/2566486.2567973
- Larsen BB, Miller EC, Rhodes MK, Wiens JJ (2017) Inordinate fondness multiplied and redistributed: The number of species on Earth and the new pie of life. The Quarterly Review of Biology 92 (3): 229-265. https://doi.org/10.1086/693564
- Marshall SA (2012) Flies: The Natural History & Diversity of Diptera. Firefly Books, Richmond Hill, Ont., Canada & Buffalo, NY, USA, 616 pp. [ISBN 978-1-77085-100-9]
- Miller E (2005) An introduction to the Resource Description Framework. Bulletin of the American Society for Information Science and Technology 25 (1): 15-19. https://doi.org/10.1002/bult.105
- Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA (2012) Uberon, an integrative multi-species anatomy ontology. Genome Biology 13 (1): R5. https://doi.org/10.1186/gb-2012-13-1-r5

- Nguyen V, Bodenreider O, Sheth A (2014) Don't like RDF reification?: Making statements about statements using singleton property. Proceedings of the 23rd international conference on World Wide Web - WWW '14. 759-770 pp. https://doi.org/10.1145/2566486.2567973
- Page L, MacFadden B, Fortes J, Soltis P, Riccardi G (2015) Digitization of biodiversity collections reveals biggest data on biodiversity. BioScience 65 (9): 841-842. https://doi.org/10.1093/biosci/biv104
- Patterson D, Mozzherin D, Shorthouse D, Thessen A (2016) Challenges with using names to link digital biodiversity information. Biodiversity Data Journal 4: e8080. https://doi.org/10.3897/bdj.4.e8080
- Poelen J, Simons J, Mungall C (2014) Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. Ecological Informatics 24: 148-159. https://doi.org/10.1016/j.ecoinf.2014.08.005
- Pyle R (2016) Towards a Global Names Architecture: The future of indexing scientific names. ZooKeys 550: 261-281. https://doi.org/10.3897/zookeys.550.10009
- Rainford JL, Mayhew PJ (2015) Diet evolution and clade richness in Hexapoda: A phylogenetic study of higher taxa. The American Naturalist 186 (6): 777-791. https://doi.org/10.1086/683461
- Remsen D (2016) The use and limits of scientific names in biological informatics. ZooKeys 550: 207-223. https://doi.org/10.3897/zookeys.550.9546
- Ren Y, Parvizi A, Mellish C, Pan JZ, Deemter Kv, Stevens R (2014) Towards competency question-driven ontology authoring. The Semantic Web: Trends and Challenges. ESWC 2014. Lecture Notes in Computer Science. Vol. 8465. Springer, Cham. https://doi.org/10.1007/978-3-319-07443-6_50
- Simsion G (2007) Data Modeling: Theory and Practice. Technics Publications, Bradley Beach, NJ.
- Simsion GC, Witt GC (2005) Data Modeling Essentials. 3rd Edition. Morgan Kaufmann Publishers, Amsterdam & Boston.
- Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C (2005) Relations in biomedical ontologies. Genome Biology 6 (5): R46.1-R46.15. https://doi.org/10.1186/gb-2005-6-5-r46
- Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Consortium TO, 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
- Walls RL, Deck J, Guralnick R, Baskauf S, Beaman R, Blum S, Bowers S, Buttigieg PL, Davies N, Endresen D, Gandolfo MA, Hanner R, Janning A, Krishtalka L, Matsunaga A, Midford P, Morrison N, Tuama ÉÓ, Schildhauer M, Smith B, Stucky BJ, Thomer A, Wieczorek J, Whitacre J, Wooley J (2014) Semantics in support of biodiversity knowledge discovery: an introduction to the Biological Collections Ontology and related ontologies. PLOS One 9 (3): e89606. https://doi.org/10.1371/journal.pone.0089606
- Walls RL, Buttigieg PL, Deck J, Guralnick R, Wieczorek J (2018) Integrating and managing biodiversity data with the Biocollections Ontology. Application of Semantic Technology in Biodiversity Science. 33. https://doi.org/10.3233/978-1-61499-854-9-81

- 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): e29715. https://doi.org/10.1371/journal.pone.0029715
- Zermoglio P, Guralnick R, Wieczorek J (2016) A Standardized Reference Data Set for Vertebrate Taxon Name Resolution. PLOS ONE 11 (1): e0146894. https://doi.org/10.1371/journal.pone.0146894

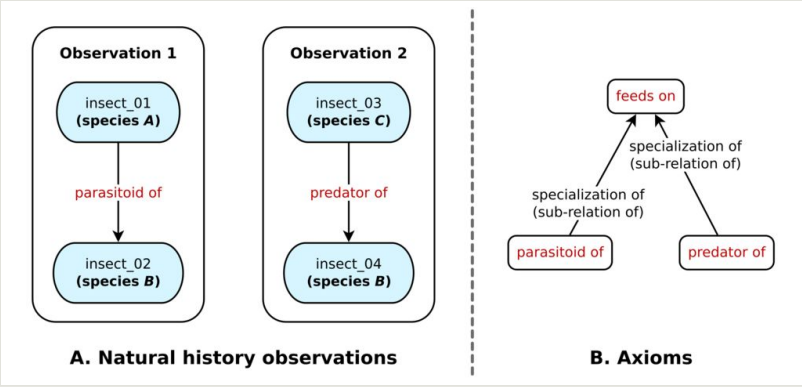

Figure 1.

Example natural history observations **A** and ontology axioms **B**. The axioms allow a computer to understand, for example, that if insect_01 is a “parasitoid of” insect_02, it is also true that insect_01 “feeds on” insect_02.

Table 1.

Ten top-level conceptual areas required for a comprehensive ontology of insect natural history data. "Relevant extant ontologies" are existing ontologies that provide at least partial coverage of the concepts in a given conceptual area. All of the ontologies mentioned here are part of the Open Biological and Biomedical Ontology (OBO) Foundry (Smith et al. 2007), a collection of interoperable ontologies that has been widely adopted in the biological sciences. Ontology references: **1.** Walls et al. (2018), **2.** Walls et al. (2014), **3.** Gene Ontology Consortium (2004), **4.** Ashburner et al. (2000), **5.** Smith et al. (2005), **6.** Gkoutos et al. (2012), **7.** Mungall et al. (2012), **8.** Buttigieg et al. (2016), **9.** Buttigieg et al. (2013), **10.** Ashburner and Schriml (2013), **11.** Dahdul et al. (2014).

Conceptual area	**Description**	**R**
Observations and observing processes	Observations of insect natural history and the processes that generate them, including information about the observers (whether human or machine) and where and when observations are made.	B
Relationships and interactions	Behaviors that involve interactions among organisms. Includes pairwise interactions (e.g., mating or herbivory) and multi-way interactions (e.g., cooperative colony defense or ants defending aphids from a potential predator).	G
Single-organism behaviors	Behaviors that do not necessarily involve interactions with other organisms (e.g., perching or locomotion).	N
Ontogeny	Developmental information (e.g., instar number or length of larval stage).	G U
Organism products and traces	Non-living objects or artifacts generated by insects (e.g., nests or leaf mines).	
Habitat, locality, and substrates	The physical context in which an organism is found, at all scales (e.g., a geopolitical boundary or a specific microhabitat).	E
Positional and spatial information	Information about the location of an organism relative to some other object or reference point (e.g., underneath the bark of a log, the south side of a rock).	B C
Weather and climate	Information about weather conditions or climate (e.g., momentary or long-term observations of temperature or precipitation) at any spatial scale.	
Collecting methods	The methods used to obtain specimens or individuals for observation (e.g., sweep netting or pitfall trapping) and information about how those methods are implemented.	B
Curation	Information about how specimens or other artifacts are managed (e.g., where they are housed and how they are preserved).	B

Supplementary materials

Suppl. material 1: Example insect natural history data (PDF document)

Authors: Brian Stucky, James Balhoff, Narayani Barve, Vijay Barve, Laura Brenskelle, Matthew H. Brush, Gregory Dahlem, James Gilbert, Akito Kawahara, Oliver Keller, Andrea Lucky, Peter Mayhew, David Plotkin, Katja Seltmann, Elijah Talamas, Gaurav Vaidya, Ramona Walls, Matt Yoder, Guanyang Zhang, Rob Guralnick
Data type: natural history
Filename: example_data.pdf - Download file (431.91 kb)

Suppl. material 2: Example insect natural history data (CSV file)

Authors: Brian Stucky, James Balhoff, Narayani Barve, Vijay Barve, Laura Brenskelle, Matthew H. Brush, Gregory Dahlem, James Gilbert, Akito Kawahara, Oliver Keller, Andrea Lucky, Peter Mayhew, David Plotkin, Katja Seltmann, Elijah Talamas, Gaurav Vaidya, Ramona Walls, Matt Yoder, Guanyang Zhang, Rob Guralnick
Data type: natural history
Filename: example_data.csv - Download file (103.95 kb)

Suppl. material 3: Ontology competency questions, user domains or groups, and example use cases

Authors: Brian Stucky, James Balhoff, Narayani Barve, Vijay Barve, Laura Brenskelle, Matthew H. Brush, Gregory Dahlem, James Gilbert, Akito Kawahara, Oliver Keller, Andrea Lucky, Peter Mayhew, David Plotkin, Katja Seltmann, Elijah Talamas, Gaurav Vaidya, Ramona Walls, Matt Yoder, Guanyang Zhang, Rob Guralnick
Data type: tables
Filename: supplemental_3.pdf - Download file (100.94 kb)