# What do Biodiversity Scholars Search for? Identifying High-Level Entities for Biological Metadata

Felicitas Löffler[1], Claas-Thido Pfaff[2], Naouel Karam[3], David Fichtmüller[4], and Friederike Klan[1]

[1] Heinz-Nixdorf Endowed Chair for Distributed Information Systems, FSU Jena, Germany
{felicitas.loeffler,friederike.klan}@uni-jena.de
[2] Systematic Botany and Functional Biodiversity Lab, University of Leipzig, Germany
{claas-thido.pfaff@uni-leipzig.de}
[3] Institute of Computer Science, Freie Universität Berlin, Germany
{naouel.karam@fu-berlin.de}
[4] Botanic Garden and Botanical Museum (BGBM), Freie Universität Berlin, Germany
{d.fichtmueller@bgbm.de}

**Abstract.** Research questions in biodiversity are as diverse and heterogeneous as data are. Most metadata standards are mainly data-focused and pay little attention to the search perspective. In this work, we introduce a method to analyze the actual information need of biodiversity scholars based on two individual studies: (1) a series of workshops with domain experts and (2) an analysis of research and search questions collected in three different biodiversity projects. We finally present 12 high-level entities that appear in all kinds of biological data across the different sources evaluated.

**Keywords:** biological data, life sciences, biodiversity, metadata, information retrieval

## 1 Introduction

In the last decade, we have witnessed an unprecedented increase of open data ranging from species-related observations, digitized specimen collections to genome or environmental data offered, e.g., through remote sensing. This opens up unforeseen opportunities particularly for biodiversity research, which relies on cross-disciplinary data analysis to elucidate the interplay between individuals and the conditions of the environment they inhabit, both on the macroscopic and microscopic level. At the flip side of this development, discovering and filtering these large volumes of multidisciplinary data becomes a more and more time-consuming and demanding task [2]. Thus, there are two big challenges: exploring effective retrieval mechanisms that support humans in finding relevant data and creating proper and rich metadata in order to make data findable (FAIR principles [6]).

The biodiversity community has responded to the latter requirement by developing metadata standards for biological data, such as Darwin Core (DwC)[5], ABCD[6] or EML[7] .

[5] Darwin Core, http://rs.tdwg.org/dwc/

[6] ABCD, http://www.tdwg.org/activities/abcd/

[7] EML, http://www.dcc.ac.uk/resources/metadata-standards/eml-ecological-metadata-language

At the same time, considerable effort has been put on the formalization of domain knowledge in terms of vocabularies and ontologies. By referencing this formal knowledge, data can be richly annotated and become machine-readable. In the last years, numerous ontologies for specific biological domains have been created, e.g., the Gene Ontology (GO)[8] for genes, the Chemical Entities of Biological Interest (ChEBI) ontology for chemical compounds, the Environmental Ontology (ENVO) [9] for environmental features and materials, the Phenotype Quality Ontology (PATO) for phenotypes and the NCBI Taxonomy[10] for species. In addition, high-level ontologies with an emphasis on interlinking biological data from different sources have been developed, e.g., the Biological Collections Ontology (BCO) [5], the Extensible Observation Ontology (OBOE) [3] or the Semantic Sensor Network Ontology (SSN) [1].

In the search applications we are hosting within the biodiversity projects GFBio (The German Federation for Biological Data)[11] and AquaDiva[12], we observe that existing metadata standards and ontologies often take a data-centric view. They provide means to well-described biodiversity data, their characteristics, their origin and the process of their creation. However, when searching for data, scholars often do not have specific data in mind, but rather a research question they would like to answer. Hence, we argue that when designing metadata standards and ontologies for biodiversity both perspectives have to be considered, the requirements given by available datasets and the way scholars are looking for data. The contribution of the paper is twofold: (1) We propose and apply a method that combines the findings of two different and independent approaches to identify high-level entities that are relevant for biodiversity researchers when searching for data (Sects. 2.1 and 2.2). (2) As a first result, we present the findings of the two individual approaches and propose a consolidated set of biological entities (Sect. 2.3). We consider this as a first step towards enriched metadata with information that is relevant to information seekers. It also serves as a prerequisite for increasing the findability of biodiversity data.

## 2 Methodology

Our approach to analyze the search perspective comprises two independent studies: Assuming that properly described data can be found more easily, the goal in dedicated workshops with scholars was to define an annotation schema that can be used to richly describe ecological data (Sect. 2.1). The second study is oriented to evaluation methods in information retrieval and analyzes research and search questions collected in three biodiversity projects (Sect. 2.2). In both approaches, the aim was to enhance search applications and to detect high-level entities that can be either used as metadata fields or that can be linked with ontologies. The first result of biological high-level entities is presented in Sect. 2.3.

[8] GO, http://www.geneontology.org/
[9] ENVO, https://github.com/EnvironmentOntology/envo
[10] NCBI Taxonomy, https://www.ncbi.nlm.nih.gov/taxonomy
[11] GFBio, https://www.gfbio.org
[12] AquaDiva, http://www.aquadiva.uni-jena.de

### 2.1 Workshops with domain experts

In close collaboration between GFBio and the German Centre for Integrative Biodiversity Research Halle – Jena – Leipzig (iDiv)[13], we conducted ten workshops with 35 domain experts from ecology and adjacent disciplines to develop a metadata schema and a controlled vocabulary, the *Essential Annotation Schema for Ecology* (EASE)[14]. This annotation framework aims at describing ecological data from a scholar's search perspective. Annotation in this context refers to metadata.

Two design principles have been formulated for the development: *Parsimony:* The framework aims at being as simple as possible in structure and content. Optimization here has to be done carefully to maintain a differentiated and consistent annotation. One example: Larger time frames in ecology are referred to by a relative reference, (e.g., 18 million years ago) or by named geological time periods. These periods are getting more granular from eons to ages and are nested in each other. It could be argued to make ages optional in the annotation which sacrifices some granularity but still maintains a consistent larger temporal context. *Comprehensiveness:* The framework aims at a certain comprehensiveness defining essential orthogonal dimensions of information which allow ecological content to be described and located in the search space of ecology. Comprehensiveness is not accomplished by using many different dimensions and concepts but rather a few essential and complementary ones which also reflect the mindset and questions of researchers when looking for data.

Based on these guidelines, 8 top level categories have been selected. During the workshops, the top level categories were substantiated in a top down approach with increasing detail (~1600 concepts). Here, we relied on expert knowledge of the contributors but also on other sources such as EML, ABCD and DwC, various topic specific textbooks (e.g., related to organic and inorganic chemistry) and standardized vocabularies (e.g., the World Reference Base for Soil[15], and The International Chronostratigraphic Chart[16]).

The top level categories are 1. *Time* (e.g., date, time, timezone), 2. *Space* (e.g., bounding box, coordinates, location names), 3. *Sphere* (e.g., pedo-, hydro-, atmosphere aspects), 4. *Biome* (e.g., zones, water availability, land use), 5. *Organism* (species classification), 6. *Process* (e.g., processes, objects and interactions), 7. *Method* (general approach, setup of gradients), 8. *Chemical* (e.g., elements, compounds, functions). In addition, the framework covers a set of general information to handle associations between primary data and annotation (e.g., data format, contact person, download URL).

### 2.2 Research and search questions in the biodiversity domain

In information retrieval, a lot of research has been done towards a perfect ranking [4] whereas little attention has been paid to a user's actual information need. What research questions are biodiversity scholars working on? What kind of data do they want to reuse? Do the provided metadata actually reflect a researcher's information need? Therefore,

[13] iDiv, https://www.idiv.de/

[14] EASE: https://github.com/cpfaff/ease

[15] WRB, http://www.fao.org/soils-portal/soil-survey/soil-classification/world-reference-base/en/

[16] ICS, http://www.stratigraphy.org/index.php/ics-chart-timescale

Table 1: Example questions gathered in three biodiversity projects

| gfbio | AquaDiva | iDiv |
|---|---|---|
| Do butterflies occur on calcareous grassland? | How does agriculture affect the groundwater composition? | How old does Plantago lanceolata get? |
| Is there data on the influence of geographic elevation on the growth rate and plant development of Zea mays? | What are suitable methods to characterize microbial soil processes by gas analytical techniques? | Do cities harbour a higher biodiversity compared to agricultural areas? |

we collected 184 search and research questions from scholars who are involved in three biodiversity projects in Germany: GFBio (73), AquaDiva (98) and iDiv (13). Examples are presented in Table 1. We asked for full questions as well as keywords to get the actual information need together with the search query. We left it to the scholars to either provide search questions posed to a search interface or broader research questions they are currently working on to get a wider spectrum of information needs.

We analyzed the questions manually and explored whether the noun entities could be grouped into high-level categories, such as *Organism* or *Environment*. For instance, given the question: *Is there DNA data about Amphimonhystrella (Nematoda)?* the noun entities are 'DNA data' and 'Amphimonhystrella (Nematoda)'. The latter one is an *Organism* whereas 'DNA data' points to a certain *Data Type*. Finally, we grouped the noun entities into 13 categories presented in Table 2. *Organism* comprises all individual life forms including plants, fungi, bacteria, animals and microorganisms. All species live in certain *Environments* and have certain characteristics that are summarized with *Quality and Phenotype*. Biological, chemical and physical *Processes* are re-occurring and transform materials or organisms due to chemical reactions or other influencing factors. *Events* are processes that appear only once at a specific time, such as environmental disasters. Chemical compounds, rocks, sand and sediments can be grouped as *Materials and Substances*. *Anatomical Entities* comprise the structure of organisms, e.g., body or plant parts, organs, cells and genes. The term *Method* describe all operations and experiments that have to be conducted to lead to a certain result. Outcomes of research methods are delivered in *Data Types*. All kinds of geographic information is summarized with *Location* and time data including geological eras are described with *Time*. *Person and Organization* are either projects or authors of data. As reflected in the search questions, scholars in biodiversity are highly interested in *Human Intervention* on landscape and environment, e.g., fishery, agriculture.

### 2.3 Discussion

Table 3 constitutes a consolidation of the main entities identified by the previously described processes. While there is a broad consensus on entities such as *Organism, Process, Method and Time*, some wording and classification on others are different. *Space* in EASE actually means location information and *Sphere* comprises altitude indications that is covered under *Environment* in the search questions. In EASE, *Biome*

Table 2: Categories selected from search questions with examples below

| | | | | |
|---|---|---|---|---|
| **Organism** | **Environment** | **Quality and Phenotype** | **Process** | **Event** |
| quercus, cyclothone, globigerina bulloides | below 4000m, ground water, city | length, growth rate, reproduction rate, traits | climate change, nitrogen transformation | Deepwater Horizon oil spill, 'Tree of the Year 2016' |
| **Material and Substance** | **Anatomical Entity** | **Method** | **Data Type** | **Location** |
| sediment, rock, CO2 | DNA, proteome, root | lidar measurements, observation, remote sensing | lidar data, sequence data | Germany, Atlantic Ocean |
| **Time** | **Person and Organization** | **Human Intervention** | | |
| current, over time, triassic | Deep Sea Drilling Project, author's or collector's names | agriculture, land use, crop yield increase | | |

contains subfields for *Land use* (*Human Intervention* in the search questions) and data attributes are defined as *Factor* under *Method* (*Quality and Phenotype* in the search questions). *Chemical* is grouped under *Material & Substance* in the search questions that additionally covers soil, sediments and rocks. *Anatomical Entity, Data Type* and *Event* only occur in the search questions.

Looking at potential linkage with existing ontologies, we finally selected 12 biological high-level entities. We left out *Sphere* since ontologies such as ENVO already cover environmental features and conditions and could be extended with altitude information. Preliminary, we will leave *Data Type* out. It needs to be further discussed and investigated whether it can be classified under other entities, e.g., *Method*.

## 3 Conclusion

We described and applied a methodology for identifying high-level entities in the biodiversity domain that reflect the scholars' point of view. In our future work, we will link the identified entities to existing ontologies. Our aim is to improve the indexing process of search applications over research data by means of these 12 categories. We would like to automatically extract information from metadata that is related to the entities. We believe, that this will help to improve data retrieval methods in the biodiversity domain.

## Acknowledgements

We would like to thank the numerous scientists from GFBio, AquaDiva and iDiv who took part in the workshops and/or provided questions. This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG) within the scope of the GFBio and AquaDiva projects.

Table 3: Consolidation of high-level entities

| **EASE** | **Questions** | **Consolidation** |
|---|---|---|
| Organism | Organism | Organism |
| Process | Process | Process |
| ✗ | Event | Event |
| Environment | Environment | Environment |
| Method - Factor | Quality & Phenotype | Quality & Phenotype |
| ✗ | Anatomical Entity | Anatomical Entity |
| Chemical | Material & Substance | Material & Substance |
| Method | Method | Method |
| ✗ | Data Type | ✗ |
| Time | Time | Time |
| Space | Location | Location |
| Sphere | ✗ | ✗ |
| General Information | Person & Organization | Person & Organization |
| Biome - Land use | Human Intervention | Human Intervention |

## References

1. M. Compton, P. Barnaghi, L. Bermudez, R. Garca-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog, V. Huang, K. Janowicz, W. D. Kelsey, D. L. Phuoc, L. Lefort, M. Leggieri, H. Neuhaus, A. Nikolov, K. Page, A. Passant, A. Sheth, and K. Taylor. The ssn ontology of the w3c semantic sensor network incubator group. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17:25 – 32, 2012.
2. M. Diepenbroek, F. Glöckner, P. Grobe, A. Güntsch, R. Huber, B. König-Ries, I. Kostadinov, J. Nieschulze, B. Seeger, R. Tolksdorf, and D. Triebel. Towards an integrated biodiversity and ecological research data management and archiving platform: GFBio. In *Informatik 2014*, 2014.
3. J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and V. F. An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, 2007.
4. C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
5. R. L. Walls, J. Deck, R. Guralnick, S. Baskauf, R. Beaman, S. Blum, S. Bowers, P. L. Buttigieg, N. Davies, D. Endresen, M. A. Gandolfo, R. Hanner, A. Janning, L. Krishtalka, A. Matsunaga, P. Midford, N. Morrison, E. O. Tuama, M. Schildhauer, B. Smith, B. J. Stucky, A. Thomer, J. Wieczorek, J. Whitacre, and J. Wooley. Semantics in support of biodiversity knowledge discovery: An introduction to the biological collections ontology and related ontologies. *PLoS ONE*, 9(3):e89606+, Mar. 2014.
6. M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. The fair guiding principles for scientific data management and stewardship. *Scientific Data 3*, (160018), 2016.