

Knowledge Graph Generation from Biodiversity Catalogs with LLM Annotation

Sefika Efeoglu^{1,2,*}, Jana Hoffmann¹

¹Senckenberg – Leibniz Institution for Biodiversity and Earth System Research, 15374 Müncheberg, Germany

²Institute of Informatics, Freie Universität Berlin, 14195 Berlin, Germany

Abstract

Type specimen (insect individual) catalogs in entomology contain large amounts of biodiversity data, including changes and losses in species numbers. Yet these catalogs lack machine readability because they are published as journal articles without annotations. Therefore, to analyze and trace how type specimens have changed in either collections or the domain over years, they must be represented in a structured format. This work constructs a domain-specific schema from competency questions (CQs) from four user groups in the entomology domain and its library, leveraging Darwin Core terms where needed. We then apply collaborative large language model annotation to the type specimen catalogs using the domain-specific schema to construct a knowledge graph. Finally, we evaluate this knowledge graph using SPARQL queries for these CQs.

Keywords

Biodiversity, Knowledge Base, Darwin Core Standards, Multimodal Scholarly Data

1. Introduction

Entomology institutes maintain type specimen catalogs that document information about insect specimens and periodically update these records. These catalogs contain valuable biodiversity information, including species names, specimen counts, collection localities (e.g., geographic coordinates and place names), and material citations, as illustrated in Figure 1. Although these catalogs are typically published in textual form, many remain only partially digitized and often contain outdated locality names, limiting their interoperability and reuse.

alluaudi LHOSTE, 1938 (*Scydmaenus*)
(Arb. morph. tax. Ent. 5, 119–120, Fig. 15)
Holotypus, Paratypen, Madagascar, Diégo-Suarez, 1893, leg. ALLUAUD; Fort Dauphin, 1900, leg. ALLUAUD. Holotypus im Mus. Paris, Paratypen im DEI.
DEI: 1 Paratypus, Tananarive von FRANZ als ssp. *tananarivensis* beschrieben (Faune de Madagaskar, im Druck); 1 Paratypus, Madagascar, F. Dauphin, 1900, leg. ALLUAUD nach FRANZ „eine andere Art oder zumindest andere ssp.“.

Figure 1: A sample of type specimen records¹.

Recent advances in large language models (LLMs) have enabled automatic generation of machine-readable knowledge from unstructured text. For instance, OntoGPT [1] demonstrates promising performance for extracting structured knowledge in high-stakes domains. However, OntoGPT is limited to entities and concepts defined in existing ontologies, making it difficult to adapt to biodiversity-specific collections and construct knowledge graphs (KGs) from the type specimen catalogs. Other LLM-based approaches, NeonGPT [2] and OntoChat [3], are applied to ontology engineering tasks, e.g., schema generation and ontology learning. Nevertheless, evaluations primarily use benchmark datasets rather than real-world data. Moreover, biodiversity data may require additional preprocessing steps, such as modernizing historical locality names. This aspect is outside the scope of the present work and remains a direction for future work.

6th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment 2026

*Corresponding author.

✉ sefika.efioglu@senckenberg.de (S. Efeoglu); jana.hoffmann@senckenberg.de (J. Hoffmann)

ORCID 0000-0002-9232-4840 (S. Efeoglu); 0000-0001-6214-661X (J. Hoffmann)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://www.contributions-to-entomology.org/article/view/1126/1125>, accessed on 21.07.2026.

This work presents a workflow for transforming the type specimen catalogs into a KG serialized in Turtle format. To design a domain-specific schema, we first conducted a workshop with four domain experts—a Lepidoptera (butterfly) researcher, the chief editor of a zoology journal, a collection manager, and the scientific head of the collection. Following a short introduction to competency questions (CQs) and user stories, the participants formulated domain-specific CQs and their user stories². These CQs guide the construction of the domain-specific schema, while reusing Darwin Core (DwC) terms where appropriate³. We then extract text from PDF catalogs using OlmoCR [4] and apply a collaborative LLM annotation process to generate triples that conform to the schema. Our main contributions in this work are:

- A CQ-driven workflow for generating the KG from the type specimen catalogs.
- A collaborative LLM-based annotation approach for the type specimen catalogs in the entomology collections.
- Releasing CQs and the user stories in the entomology domain [5].
- Evaluation of CQs on the KG of the type specimen catalog.

In the rest of this paper, Section 2 summarizes related works on KG construction from biodiversity texts. Section 3 presents the proposed workflow. Section 4 presents and discusses experimental results. Finally, Section 5 summarizes the conclusions and future work.

2. Related Work

This section presents and discusses related works on biodiversity research data management systems and metadata standards.

Several research data management systems have been developed to support semantic publishing and knowledge management in biodiversity. OpenBiodiv-O [6] provides a domain ontology for the OpenBiodiv knowledge management system, enabling biodiversity publications to be represented as machine-readable KGs while reusing DwC terms. Likewise, BiodivPortal [7], a partner of NFDI4Biodiversity and OntoPortal, offers ontology repositories, along with annotation and mapping services for biodiversity data. Both platforms primarily rely on existing ontologies and controlled vocabularies for semantic annotation.

Biodiversity metadata standards provide the foundation for interoperable knowledge representation. Agosti et al. [8] discuss metadata requirements for biodiversity publications, emphasizing persistent identifiers, material citations, and FAIR (Findable, Accessible, Interoperable, and Reusable) principles. LifeWatch⁴ [9] offers an infrastructure for managing biodiversity research data and supports the publication and curation of biological collections and associated media. Darwin Core (DwC) [10], developed by the Biodiversity Information Standards (TDWG) community, has become the de facto standard for describing biodiversity occurrence and specimen data.

Our work complements these efforts by combining competency question-driven schema generation with collaborative, LLM-based annotation to generate a KG from type specimen catalogs. The resulting workflow automatically transforms the type specimen catalogs into the KG that supports semantic querying and integration with biodiversity information systems.

3. Methodology

We propose a collaborative LLM annotation workflow for constructing KGs from type specimen catalogs. The workflow combines a domain-specific schema derived from competency questions (CQs) with collaborative LLM annotation of catalog records. As illustrated in Figure 2, the workflow consists of four

²Available at https://github.com/Senckenberg-Nature-Research-SDEI/collaborative_text_annotation/tree/master/doc/user-scenarios, accessed on 21.07.2026.

³<https://dwc.tdwg.org/list/>, accessed on 24.07.2026.

⁴<https://metadatalogue.lifewatch.eu/srv/eng/catalog.search#/home>, accessed on 21.07.2026.

stages: (i) competency questions, (ii) CQs to schema (Turtle), (iii) manually refining the schema, and (iv) schema enrichment with individuals extracted from the type specimen catalogs. A type specimen KG is generated at the end of the workflow in Figure 2.

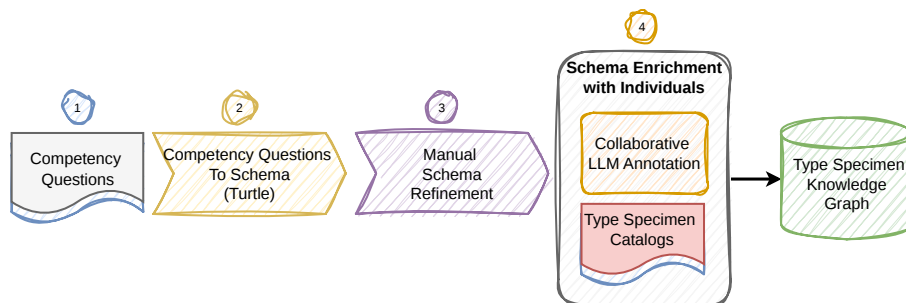


Figure 2: A collaborative LLM annotation workflow for generating knowledge graphs from type specimen catalogs.

3.1. Schema Generation from Competency Questions

We organized a workshop in which four domain experts—a Lepidoptera (butterfly) researcher, the chief editor of a zoology journal, a collection manager, and the scientific head of the collection—formulated user stories and CQs following a short introduction about how CQs and user stories can be written (see footnote 2). We then used the resulting CQs to derive an initial schema with an LLM, which extracts candidate classes and properties and serializes them into Turtle format. We then refined the generated schema in Protégé following “Ontology Development 101” in [11]. The authors performed the manual refinement without direct involvement of the domain experts, although one author is an expert in the entomology domain. The CQs and the design principles described below guided the refinement decisions. For instance, this manual refinement process included changing the hierarchy of classes and properties, adding missing classes and properties stated in the CQs, and removing duplicate annotations. In the schema generation stage, we applied the following design principles:

- Reuse Darwin Core classes and properties whenever applicable.
- Specify domain and range using DwC classes whenever possible.
- Keep class and property names concise and singular.
- Model only information explicitly implied by the competency questions.
- Avoid introducing concepts that are not supported by the domain requirements.

3.2. Schema Enrichment with Individuals

We first convert the type specimen catalogs from PDF to text using OlmoCR [4], an LLM-based optical character recognition model. Since each catalog contains numerous specimen records (see Figure 1), the OCR-extracted text was first manually cleaned and then manually split into individual specimen records before annotation. A group of LLMs then processes each record, using the generated domain-specific schema as guidance. Collaborative LLM annotation refers to an autonomous, multi-model process in which multiple LLMs independently annotate the same record. The system aggregates their outputs using a voting-based aggregation procedure. Thus, collaboration refers to aggregating multiple LLM outputs rather than human–AI collaboration. Every model independently extracted resources (entities) and their relations, producing triples that conform to the schema. The final annotation is determined by aggregating the LLM outputs. A triple is retained if at least two of the four LLMs produce the same triple.

4. Evaluation

This section presents experimental details in Section 4.1. Then, Section 4.2 presents and discusses the results.

4.1. Experimental Settings

- **LLMs for collaborative annotation.** We used Qwen3-8B, Llama-3.1-8B-Instruct, DeepSeek-R1-Distill-Qwen-7B, and Mistral-7B-Instruct-v0.3. We selected these models without a comparative model-selection evaluation.
- **Schema generation with competency questions.** GPT-4.1-mini is used.
- **Dataset:** 1,032 catalog records from [12].
- **Competency question-based evaluation.** The domain-specific KG was evaluated using SPARQL versions⁵ of CQs. We then checked whether the results matched the personas’ requests. Listing 1 presents a SPARQL version of a CQ written by the butterfly researcher.
- **Hardware:** The experiments were conducted on Google Colab Pro using an NVIDIA A100 GPU.

```
1 SELECT ?specimen ?publication ?title ?taxon ?institution ?locality ?identifier
2 WHERE {
3   ?publication ex:documentsTypeSpecimen ?specimen .
4
5   OPTIONAL { ?publication dcterms:title ?title . }
6   OPTIONAL { ?specimen dwc:scientificName ?taxon . }
7   OPTIONAL { ?specimen ex:heldBy ?institution . }
8   OPTIONAL { ?specimen dwc:locality ?locality . }
9   OPTIONAL { ?specimen dcterms:identifier ?identifier . }
10 }
```

Listing 1: The SPARQL version of “Which preserved specimens are held by a specified institution?” written by a Lepidoptera (butterfly) researcher.

4.2. Results and Discussion

We applied our workflow to the published type specimen catalog in [12] to transform it into Turtle format. An example specimen record from the original catalog is shown in Figure 1, while its corresponding Turtle representation is presented in Listing 2. We then executed the SPARQL implementations of the CQs on the generated dataset in Turtle format. The execution results, including the number of records returned by each query, are summarized in Table 1.

The results show that Turtle effectively answers CQs that rely on structured information in a specimen catalog. In particular, the *Chief Editor* and *Collection Manager* personas can answer most of their CQs using data from a single catalog. Likewise, the *Scientific Head* can answer seven of the eleven CQs.

In contrast, several CQs defined for the *Butterfly Researcher* cannot be answered using the current dataset alone. These include questions that require external knowledge enrichment, such as taxonomic synonymy, type specimen repositories, and links to associated digital resources. In addition, the CQ concerning temporal trends in species descriptions requires aggregating information across multiple specimen catalogs and therefore cannot be answered from a single dataset. Such questions require either integration with external knowledge sources or application-level processing.

⁵https://github.com/Senckenberg-Nature-Research-SDEI/collaborative_text_annotation/tree/master/sparqls, accessed on 21.07.2026.

Table 1: Results for the SPARQL queries of competency questions (CQs) on specimen catalogs. Each record includes more than one type specimen (insect individual) about scientific species.

User Role	Competency Query	Rows
Butterfly Researcher	CQ1 – Taxonomic synonymy	0
Butterfly Researcher	CQ2 – Type specimen repository	0
Butterfly Researcher	CQ3 – Associated digital resources	0
Butterfly Researcher	CQ4 – Temporal trends in species descriptions	0
Butterfly Researcher	CQ5 – Type locality information	210
Butterfly Researcher	CQ6 – Geographic distribution of type specimens	85
Butterfly Researcher	CQ7 – Taxonomic authorship	208
Butterfly Researcher	CQ8 – Collection holdings	7
Chief Editor	CQ1 – Published type specimens and metadata	210
Chief Editor	CQ2 – Type specimen quantification	1
Chief Editor	CQ3 – Collection and repository holdings	6
Chief Editor	CQ4 – Taxonomic distribution	97
Chief Editor	CQ5 – Geographic distribution	85
Chief Editor	CQ6 – Relevant publications and catalogues	0
Chief Editor	CQ7 – Publication context	210
Chief Editor	CQ8 – Major contributors	53
Chief Editor	CQ9 – Temporal trends	30
Collection Manager	CQ1 – Type specimen documentation	210
Collection Manager	CQ2 – Available catalogues and publications	0
Collection Manager	CQ3 – Specialist contributions	53
Collection Manager	CQ4 – Relevant publications	99
Collection Manager	CQ5 – Metadata availability	210
Collection Manager	CQ6 – Documentation completeness	210
Collection Manager	CQ7 – Catalogue coverage	0
Collection Manager	CQ8 – Multiple documentation sources	186
Collection Manager	CQ9 – Historical changes in documentation	3850
Collection Manager	CQ10 – Identifiers and external resources	210
Collection Manager	CQ11 – Richly connected specimens	210
Collection Manager	CQ12 – Information diversity and coverage	0
Scientific Head	CQ1 – Type specimen inventory	210
Scientific Head	CQ2 – Type specimen quantification	1
Scientific Head	CQ3 – Taxonomic distribution	74
Scientific Head	CQ4 – Collection strengths by taxon	1
Scientific Head	CQ5 – Geographic coverage	33
Scientific Head	CQ6 – Published catalogues	0
Scientific Head	CQ7 – Specialist contributions	0
Scientific Head	CQ8 – Historical interpretation	0
Scientific Head	CQ9 – Collection development	30
Scientific Head	CQ10 – Collection strength assessment	1
Scientific Head	CQ11 – Knowledge contribution of catalogues	0

```

dwc:bibliographicCitation "Revue Suisse de Zoologie, Genève 21: 671-672" ;
dwc:country "Cameroon" ;
dwc:genus "Camponotus" ;
dwc:individualCount 1 ;
dwc:institutionCode "DEI" ;
dwc:recordedBy "Conradt" ;
dwc:scientificName "Camponotus ferreri" ;
dwc:scientificNameAuthorship "Forel, 1913" ;
dwc:sex "female" ;
dwc:specificEpithet "ferreri" ;
dwc:typeStatus "Holotype" ;
dwc:verbatimLabel "Holotypus, Kamerun, leg. Conradt, im DEI." ;
dwc:verbatimLocality "Kamerun" ;
sdei_kb:catalogueRecord ""ferreri Forel, 1913 (Camponotus)
Revue Suisse de Zoologie, Genève 21: 671-672
Holotypus, Kamerun, leg. Conradt, im DEI."" ;

```

Listing 2: Example specimen record

In summary, the generated dataset in Turtle format successfully supports quantitative and descriptive queries over specimen metadata. However, the CQs that depend on semantic enrichment, cross-catalog integration, or historical analyses remain outside the scope of the current dataset. Furthermore, because we evaluated a single specimen catalog, the *Collection Manager*'s CQ regarding available catalogs and publications (CQ2) could not be fully addressed. Extending the workflow to process multiple catalogs and integrating additional external resources are important directions for future work. In addition, domain experts did not independently validate the LLM annotations. Such domain-expert validation remains future work.

5. Conclusion

We presented a collaborative LLM annotation workflow for transforming type specimen catalogs into KGs using a domain-specific schema derived from competency questions (CQs). The proposed workflow converts PDF catalogs into machine-readable, searchable KGs that can be queried using SPARQL. Furthermore, the generated KGs can be exported as a spreadsheet using DwC terms, facilitating integration with biodiversity information systems, e.g., the Digital Information System for Natural History Data (DINA)⁶ (see Figure 3 in the Appendix for an example displayed in DINA). Furthermore, we release a collection of CQs developed by four domain experts, providing a reusable resource for future research on ontology engineering and LLM-assisted KG construction.

The current workflow has limitations. First, its performance depends on the quality of the Optical Character Recognition (OCR) output, as many historical type specimen catalogs are available only as scanned PDFs. OCR errors may therefore propagate to the downstream annotation process. Second, we elicit CQs from only four stakeholder personas within the entomology domain, which may miss the full range of user requirements across different taxonomic ranks (orders).

As future work, we plan to incorporate additional domain experts representing other insect orders to broaden the CQ collection and improve schema coverage. We also intend to extend the workflow to additional biodiversity catalog types, e.g., genus catalogs and other taxonomic collections, while investigating more robust OCR and collaborative LLM annotation strategies to improve the quality of the generated KGs. Domain-expert validation remains future work.

Declaration of Generative AI Use

During the preparation of this work, a generative AI model (GPT 5.1 with a temporary chat option) was used to check grammar and spelling. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

⁶<https://www.dina-project.net/de/consortium/about>

References

- [1] J. H. Caufield, H. Hegde, V. Emonet, N. L. Harris, M. P. Joachimiak, N. Matentzoglou, H. Kim, S. Moxon, J. T. Reese, M. A. Haendel, P. N. Robinson, C. J. Mungall, Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES): a method for populating knowledge bases using zero-shot learning, *Bioinformatics* 40 (2024) btae104. URL: <https://doi.org/10.1093/bioinformatics/btae104>. doi:10.1093/bioinformatics/btae104. arXiv:<https://academic.oup.com/bioinformatics/article-pdf/40/3/btae104/56912793/btae104.pdf>.
- [2] N. Fathallah, A. Das, S. D. Giorgis, A. Poltronieri, P. Haase, L. Kovriguina, NeOn-GPT: A Large Language Model-Powered Pipeline for Ontology Learning, in: *The Semantic Web: ESWC 2024 Satellite Events: Hersonissos, Crete, Greece, May 26–30, 2024, Proceedings, Part I*, Springer-Verlag, Berlin, Heidelberg, 2024, p. 36–50. URL: https://doi.org/10.1007/978-3-031-78952-6_4. doi:10.1007/978-3-031-78952-6_4.
- [3] B. Zhang, V. A. Carriero, K. Schreiberhuber, S. Tsaneva, L. S. González, J. Kim, J. de Berardinis, OntoChat: Framework for Conversational Ontology Engineering Using Language Models, in: *The Semantic Web: ESWC 2024 Satellite Events: Hersonissos, Crete, Greece, May 26–30, 2024, Proceedings, Part I*, Springer-Verlag, Berlin, Heidelberg, 2025, p. 102–121. URL: https://doi.org/10.1007/978-3-031-78952-6_10. doi:10.1007/978-3-031-78952-6_10.
- [4] J. Poznanski, J. Borchardt, J. Dunkelberger, R. Huff, D. Lin, A. Rangapur, C. Wilhelm, K. Lo, L. Soldaini, olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models, 2025. URL: <https://arxiv.org/abs/2502.18443>. arXiv:2502.18443.
- [5] S. Efeoglu, J. Hoffmann, User Scenarios and Competency Questions—Knowledge Graph Generation from Biodiversity Catalogs with LLM Annotation, 2026. URL: <https://doi.org/10.5281/zenodo.22794158>. doi:10.5281/zenodo.22794158, Sci-K Workshop @ ISWC 2026.
- [6] V. Senderov, K. Simov, N. Franz, P. Stoev, T. Catapano, D. Agosti, G. Sautter, R. A. Morris, L. Penev, OpenBiodiv-O: Ontology of the OpenBiodiv Knowledge Management System, *Journal of Biomedical Semantics* 9 (2018) 5. URL: <https://doi.org/10.1186/s13326-017-0174-5>. doi:10.1186/s13326-017-0174-5.
- [7] N. Karam, J. Fillies, C. Jonquet, S. Bouazzouni, F. Löffler, F. Zander, B. König-Ries, A. Güntsch, M. Diepenbroek, A. Paschke, BiodivPortal: Enabling Semantic Services for Biodiversity within the German National Research Data Infrastructure, *Datenbank-Spektrum* 24 (2024) 129–137. URL: <https://doi.org/10.1007/s13222-024-00474-5>. doi:10.1007/s13222-024-00474-5.
- [8] D. Agosti, L. Benichou, W. Addink, C. Arvanitidis, T. Catapano, G. Cochrane, M. Dillen, M. Döring, T. Georgiev, I. Gérard, Q. Groom, P. Kishor, A. Kroh, J. Kvaček, P. Mergen, D. Mietchen, J. Pauperio, G. Sautter, L. Penev, Recommendations for Use of Annotations and Persistent Identifiers in Taxonomy and Biodiversity Publishing, *Research Ideas and Outcomes* 8 (2022) e97374. URL: <https://doi.org/10.3897/rio.8.e97374>. doi:10.3897/rio.8.e97374. arXiv:<https://doi.org/10.3897/rio.8.e97374>.
- [9] L. Vaira, E. Tzafesta, M. Fiaz, H. B. Salah, A. Basset, The LifeWatch ERIC Metadata Catalogue: A FAIR gateway to biodiversity and ecosystem research, *Ecological Informatics* 96 (2026) 103853. URL: <https://www.sciencedirect.com/science/article/pii/S1574954126002591>. doi:<https://doi.org/10.1016/j.ecoinf.2026.103853>.
- [10] J. Wiczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson, D. Vieglais, E. G. Denny, R. De Giovanni, et al., Darwin Core: An Evolving Community-Developed Biodiversity Data Standard, *PLOS ONE* 7 (2012) e29715. URL: <https://doi.org/10.1371/journal.pone.0029715>. doi:10.1371/journal.pone.0029715.
- [11] N. Noy, D. McGuinness, et al., Ontology development 101: A guide to Creating Your First Ontology, 2001. URL: <http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html>.
- [12] A. Taeger, H. Gaedike, S. M. Blank, Katalog der primären Hymenopteren-Typen des DEI (unter Ausschluss der Symphyta und Sphecidae s. l.), *Beiträge zur Entomologie = Contributions to Entomology* 55 (2005) 151–250. URL: <https://www.contributions-to-entomology.org/article/view/1624>. doi:10.21248/contrib.entomol.55.1.151-250.

Appendix

staging-dina.senckenberg.de/collection/material-sample/view?id=019fd63c-fdcc-7499-a692-147034913de5

[Back to list page](#) Form Template [Edit](#) [Split](#) [Generate Label](#) [Revisions](#) [Delete](#)

Select Column

Total matched records: 0

« First < Previous Page 1 of 1 Rows Per Page 500 Next > Last »

Primary ID	Material Sample Type	Material Sample State	Created On	Tags	Actions
No Rows Found					

« First < Previous Page 1 of 1 Rows Per Page 500 Next > Last »

Total matched records: 0

Material Sample Info

Material Sample Type [Material Sample Type](#)

Material Sample State [Material Sample State](#)

Etikett [Etikett](#)

Use As Base Name For Split ☐

Material Sample Remarks

Literaturzitat [Literaturzitat](#)

Holotypus, Ceylon, Pankulam, 1899, leg. Horn, im DEL. Entomologische Mitteilungen, Berlin-Dahlem 1: 144-146

Collecting Event

[Details Page](#)

Identifiers

Collection Number [Collection Number](#)

Additional Collection Numbers

Collecting Date

Verbatim Event DateTime

1899

Start Event Date Time [Start Event Date Time](#)

End Event Date Time [End Event Date Time](#)

Collecting Agents

Verbatim Collectors

Collectors

[Horn \(legacy id: 29942\)](#)

Collector's Number [Collector's Number](#)

(a) Material sample information.

[Back to list page](#) Form Template [Edit](#) [Split](#) [Generate Label](#) [Revisions](#) [Delete](#)

Organisms

Organisms quantity

1

Life Stage

Sex

female

Organism Remarks

Vernacular Name

Grouped Individual Quantity [Grouped Individual Quantity](#)

1

Determinations

Verbatim Determination

Verbatim Scientific Name

Verbatim Determiner

Verbatim Date

Verbatim Remarks

Transcriber's Notes

Determination

Scientific Name

Leucospis viridissima

Species : viridissima > Genus : Leucospis

Determining Agents

Determining Date

Determination Remarks

Type Specimen

Type Status

Holotype

Type Status Evidence

(b) Organism information.

Figure 3: Annotated type specimen displayed in DINA: (a) Material sample information and (b) Organism information.