

# Shortening the "long tail" of "dark data": **Automated export of small data from Biodiversity Data Journal to GBIF and EOL through Darwin Core Archive**

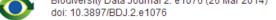
Markus Döring<sup>1</sup>, Tim Robertson<sup>1</sup>, Teodor Georgiev<sup>2</sup>, Jordan Biserkov<sup>2</sup>, Pavel Stoev<sup>2</sup>, Patrick Leary<sup>3</sup>, Katja Schulz<sup>3</sup>, Cynthia Parr<sup>3</sup>, Guido Sautter<sup>4</sup>, Donat Agosti<sup>4</sup>, Lyubomir Penev<sup>2</sup>

<sup>1</sup> GBIF, Copenhagen, Denmark <sup>2</sup> Pensoft Publishers, Sofia, Bulgaria <sup>3</sup> Encyclopedia of Life, Woods Hole, USA <sup>4</sup> Plazi, Bern, Switzerland

PENSOFT.	1	Home	About BDJ	Articles	Books	E-Books	Journals	News & Blog	Contact		<u>Register</u>   <u>Login</u>
Biodiversity Data Jo doi: 10.3897/BDJ.2	ournal 2: e103 .e1076	76 (26 Mar	2014)					209		Contents Article info Citation Metrics S	Share Review it

Download as CSV 🖾





#### Taxon treatments

Crassignatha Wunderlich, 1995

#### Nomenclature

Crassignatha Wunderlich, 1995 - Wunderlich 1995: 546; Miller et al. 2009: 68.

#### Type species

Crassignatha haeneli Wunderlich, 1995

#### Crassignatha danaugirangensis, sp. n.

ZooBank urn:lsid:zoobank.org:act:EDB4926E-0CBB-448F-B657-10373F6FD69F

#### Holotype:

Materials

a. scientificName: Crassignatha danaugirangensis; order: Araneae; family: Symphytognathidae; taxonRank: species; genus: Crassignatha; specificEpithet: danaugirangensis; scientificNameAuthorship: Miller et al. 2014; island: Borneo; country: Malaysia; stateProvince: Sabah; locality: Danau Girang Field Centre, plot 1; verbatimCoordinates: 5°24.75'N 118°2.35'E; decimalLatitude: 5.4125; decimalLongitude: 118.0392; samplingProtocol: dusting for webs; eventDate: 2014-03-04; individualCount: 1; sex: 1 male; catalogNumber: 20140304.1:57H; recordedBy: J. Miller, C.M. van der Graaf, C. Burmester; institutionID: Universiti Malaysia Sabah; collectionID: Institute for Tropical Biology and Conservation, Borneensis; institutionCode: UMS; collectionCode: BORN; basisOfRecord: PreservedSpecimen

#### Paratypes:

a. scientificName: Crassignatha danaugirangensis; order: Araneae; family: Symphytognathidae; taxonRank: species; genus: Crassignatha; specificEpithet: danaugirangensis; scientificNameAuthorship: Miller et al. 2014; island: Borneo; country: Malaysia; stateProvince: Sabah; locality: Danau Girang Field Centre, plot 1; verbatimCoordinates: 5°24.75'N 118°2.35'E; decimalLatitude: 5.4125; decimalLongitude: 118.0392; samplingProtocol: dusting for webs; eventDate: 2014-03-04; individualCount: 4; sex: 4 females; catalogNumber: 20140304.1:57; recordedBy: J. Miller, C.M. van der Graaf, C. Burmester; institutionID: Universiti Malaysia Sabah; collectionID: Institute for Tropical Biology and Conservation, Borneensis; institutionCode: UMS; collectionCode: BORN; basisOfRecord: PreservedSpecimen b. scientificName: Crassignatha danaugirangensis; order: Araneae; family: Symphytognathidae; taxonRank: species; genus: Crassignatha; specificEpithet: danaugirangensis; scientificNameAuthorship: Miller et al. 2014; island: Borneo; country: Malaysia; stateProvi Sabah; locality: Danau Girang Field Centre, plot 1; verbatimCoordinates: 5°24.75'N 118

# Pensoft Publishers PROFILE SIGN OUT GO Encyclopedia of Life Crassignatha danaugirangensis add to a collection learn more about names for this taxon Like 🗧 0 Tweet 0 Overview Detail Data 9 Media 0 Maps Names Community Resources Literature Updates EOL has no trait data No one has contributed data records for Crassignatha danaugirangensis yet. Learn how to

Figures Tables Map Taxa Data References 学 Tables and Figures, if present, can be downloaded from the article. Download all occurrences as Darwin Core Archive U Download all treatments as Darwin Core Archive Supplementary material 1 Spider morphospecies sampled from rip rian forest, riverine forest, and oil palm plantation Authors: Jeremy A. Miller, Jennie Lilliendahl Burnester, Lot van der Graaf Data type: Structured sampling data Brief description: Adult ground-web-building spider morphospecies sampled from 1 m<sup>2</sup> plots in the Danau Girang botanica plots and nearby Hilco Estate oil palm plantation. The number of 1 m<sup>2</sup> plots in each site is given as *n*. Botanical plots 1 and 4 are riparian forest habitat, plots 2 and 3 are riverine forest habitat subject to frequent inundation. Filename: morphospecies.xls Download file (23.00 kb)

# Dispatch from the field: ecology of ...

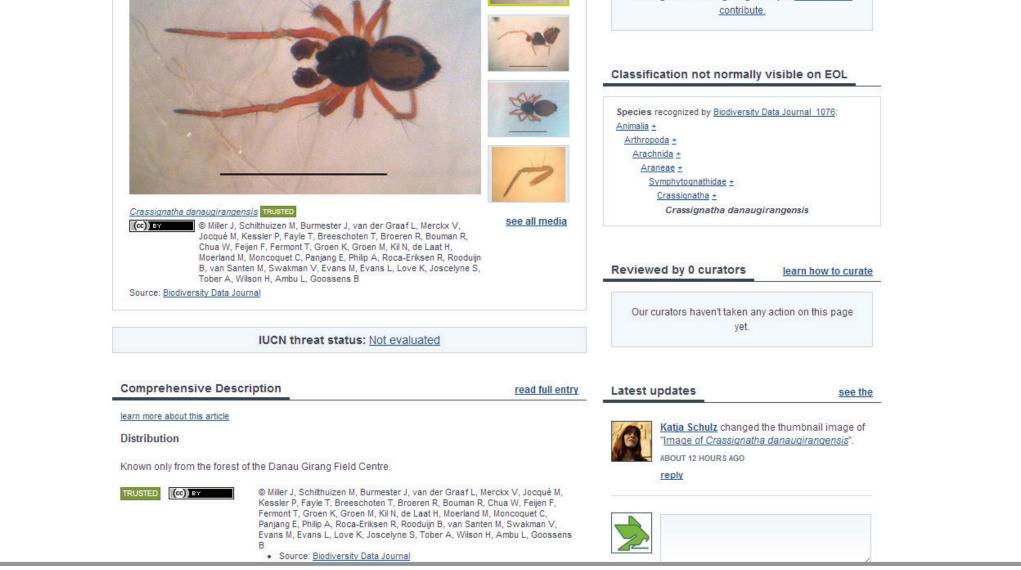
12 Occurrences View occurrences

### Summary

Information

FULL TITLE			PUBLISHED BY
Dispatch from the field: ecol species (Araneae, Symphyte		spiders with description of a new	w Biodiversity Data Journal
	ognadinado)		PUBLICATION DATE
DESCRIPTION			24-Mar-2014
		ytognathidae) was discovered Girang Field Centre in Sabah,	REGISTRATION DATE
Malaysia. A taxonomic desc	ription and accompanying e	cological study were completed which belongs to the ground-w	
building spider community, t	SERVED BY		
nundated riverine forest, and the most abundant ground-v			
rom the recently inundated	forest and was not found in a	a nearby oil palm plantation. The	e LINKS
		te the accumulation of data abo haping invertebrate communitie:	
ANGUAGE OF DATA			EXTERNAL DATA
ANGUAGE OF DATA			Darwin Core Archive
		0010114700	METADATA DOCUMENTS
ADMINISTRATIVE CONTACT Jeremy Miller	METADATA AUTHOR Jeremy Miller	ORIGINATOR Jeremy Miller	<ul> <li>GBIF annotated version (EML) </li> </ul>
	· .		
	14.		12 Georeferenced
		ſ	data
			4px data
<b>k</b>			VIEW RECORDS
<u>*</u> بر			+px
			VIEW RECORDS

According to a recent study, 80 % of scientific data are lost within two decades, at an alarming rate (Vines et al. 2014). The majority of biodiversity data is usually not indexed or properly preserved and form the so called "long tail" of "dark data", which therefore is more likely to remain invisible to scientists and eventually lost (Heidorn 2008). The bulk of "dark" biodiversity data is constituted by small and scattered datasets, especially species occurrences, published in various literature sources or in grey literature.



## Literature

Heidorn PB (2008) Shedding Light on the Dark Data in the Long Tail of Science. Library Trends 57(2) Fall 2008. Institutional Repositories: Institutional Repositories: Current State and Future. Edited by Sarah Sheeves and Melissa Cragin. (http://hdl.handle.net/2142/9127).

Miller J, Schilthuizen M, Burmester J, van der Graaf L, Merckx V, Jocqué M, Kessler P, Fayle T, Breeschoten T, Broeren R, Bouman R, Chua W, Feijen F, Fermont T, Groen K, Groen M, Kil N, de Laat H, Moerland M, Moncoquet C, Panjang E, Philip A, Roca-Eriksen R, Rooduijn B, van Santen M, Swakman V, Evans M, Evans L, Love K, Joscelyne S, Tober A, Wilson H, Ambu L, Goossens B (2014) Dispatch from the field: ecology of micro web-building spiders with description of a new species. Biodiversity Data Journal 2: e1076. DOI: 10.3897/BDJ.2.e1076

Vines TH, Albert AYK, Andrew RL, Débarre F, Bock DG, Franklin MT, Gilbert KJ, Moore J-S, Renaut S, Rennison DJ (2014) The availability of research data declines rapidly with article age. Current

The Biodiversity Data Journal uses a novel workflow that allows for upfront markup of occurrence data via its own article authoring tool, the Pensoft Writing Tool. Occurrence data, if compliant to Darwin Core, can be imported into the manuscript in a human-readable format and downloaded from the published article as a CSV file per each separate taxon treatment. In addition, all occurrence data published within an article, are automatically exported on the day of publication into a Darwin Core Archive.

The Darwin Core Archive is then registered with GBIF through their RESTful API so it becomes immediately visible and the data subsequently indexed by the GBIF data portal. The whole process happens on the day of publication, saving in this way a great deal of effort to markup occurrences and export these to aggregators. The same DwC archive is used by Encyclopedia of Life (EOL) to harvest richer data, such as descriptions, images, bibliographies, etc.

The workflow was successfully tested with the paper of Miller et al. (2014) which is remarkable also as probably the fastest ever description and peer-reviewed publication of a new species. It took less than 30 days from discovery in the field station of Malaysian Borneo to publication and data sharing in GBIF and EOL.

The elaboration of the workflow from BDJ and Plazi to GBIF and EOL through Darwin Core Archive was supported by the EU-funded project EU BON, grant agreement No 308454.

