EOL TraitBank

An open digital repository for organism traits

开放的生物性状（特征）数据仓库

Bob Corrigan

Director of Operations, Encyclopedia of Life
Smithsonian Institution
National Museum of Natural History

4th Conference on Biodiversity Informatics in China
November 4, 2015

Semantic Web -1 (2015) 1–12 1
DOI 10.3233/SW-150190
IOS Press

TraitBank: Practical semantics for organism attribute data

Editor(s): Benjamin Adams, The University of Auckland, New Zealand; Nico Franz, Arizone State University, USA; Birgitta König-Ries, Friedrich-Schiller University Jena, Germany; Deborah McGuinness, Rensselaer Polytechnic Institute, USA; Mark Schildhauer, NCEAS, University of California, Santa Barbara, USA
Solicited review(s): Jens Kattge, Max Planck Institute for Biogeochemistry, Germany; Birgitta König-Ries, Friedrich-Schiller University Jena, Germany; 2 anonymous reviewers

Cynthia S. Parr [a,*], Katja S. Schulz [a,**], Jennifer Hammock [a], Nathan Wilson [b,***], Patrick Leary [b,****], Jeremy Rice [b] and Robert J. Corrigan Jr. [a]
[a] *National Museum of Natural History, Smithsonian Institution, Washington, DC, USA*
E-mails: esparr@gmail.com, schulzk@si.edu
[b] *Marine Biological Laboratory, Woods Hole, Massachusetts, USA*

Abstract. Encyclopedia of Life (EOL) has developed TraitBank (http://eol.org/traitbank), a new repository for organism attribute (trait) data. TraitBank aggregates, manages and serves attribute data for organisms across the tree of life, including life history characteristics, habitats, distributions, ecological relationships and other data types. We describe how TraitBank ingests and manages these data in a way that leverages EOL's existing infrastructure and semantic annotations to facilitate reasoning across the TraitBank corpus and interoperability with other resources. We also discuss TraitBank's impact on users and collaborators and the challenges and benefits of our lightweight, scalable approach to the integration of biodiversity data.

Keywords: Biodiversity, ontologies, Semantic Web, traits, ecology, evolution, taxonomy, data aggregation

1. Introduction

While human knowledge of life on Earth is vast, there is no easy way to query all the information accumulated in hundreds of years of biodiversity research and documentation. Even simple questions like "which plants have yellow flowers?" or "what do sharks eat?" are impossible to answer with confidence.

Biologists have captured and managed information about morphology, behavior, life history, and ecological interactions in many different ways. Most of this information survives in the form of free text or data tables in published papers, if it survives at all [20]. Lately communities have started to annotate those papers [3], extract information from text [28,40], and build special-purpose databases of trait data, for example, TRY[1] for plants [24] and SeaLifeBase[2] for marine organisms. In addition, modern researchers are more likely to archive and share data sets associated with their published studies in open data repositories such as Dryad[3] [42], Ecological Archives[4] and PANGAEA.[5] While these are critical developments, there is still little standardization in how biologists talk about the characteristics of organisms, how they

*Current address: United States Department of Agriculture, Agricultural Research Service, Beltsville, Maryland, USA.
**Corresponding author. E-mail: schulzk@si.edu.
***Current address: MassChallenge, Boston, Massachusetts, USA.
****Current address: California Academy of Sciences, San Francisco, California, USA.

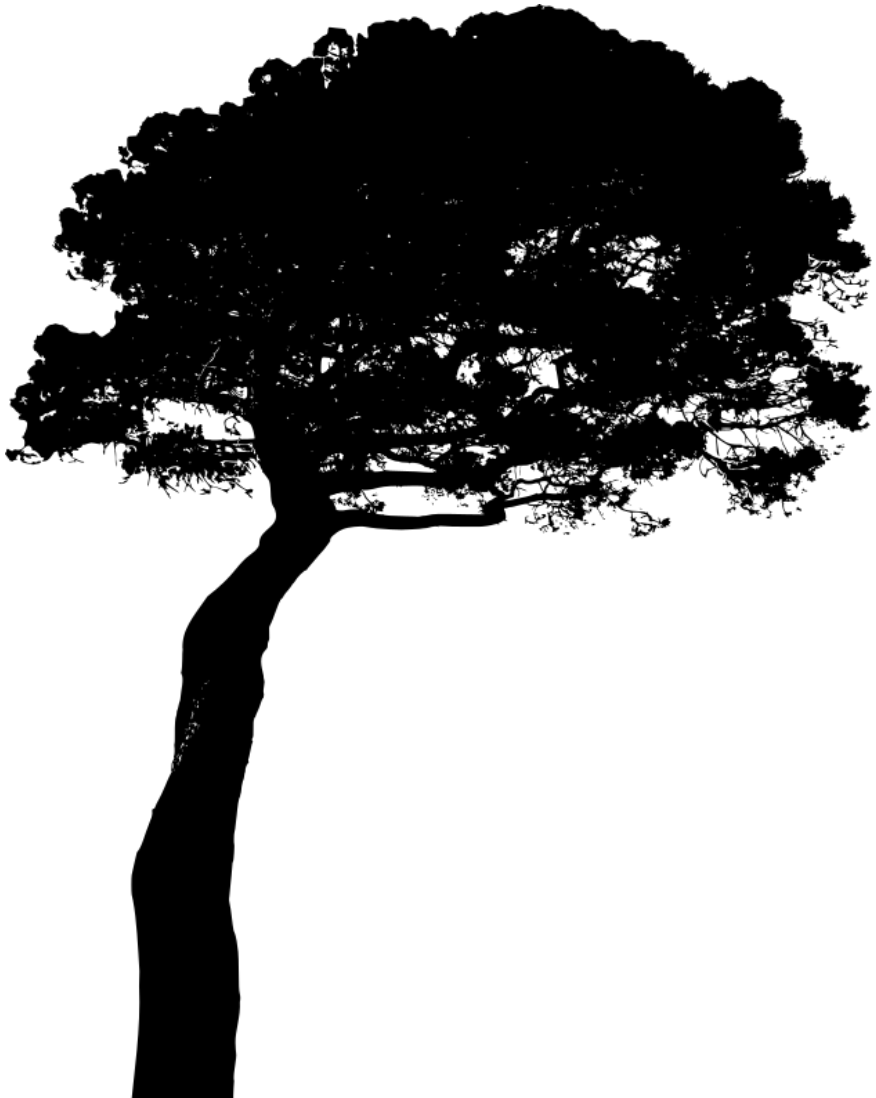
[1] http://www.try-db.org
[2] http://sealifebase.org
[3] http://datadryad.org/
[4] http://esapubs.org/archive/
[5] http://www.pangaea.de

Semantic Web Journal

In Press

DOI 10.3233/SW-150190

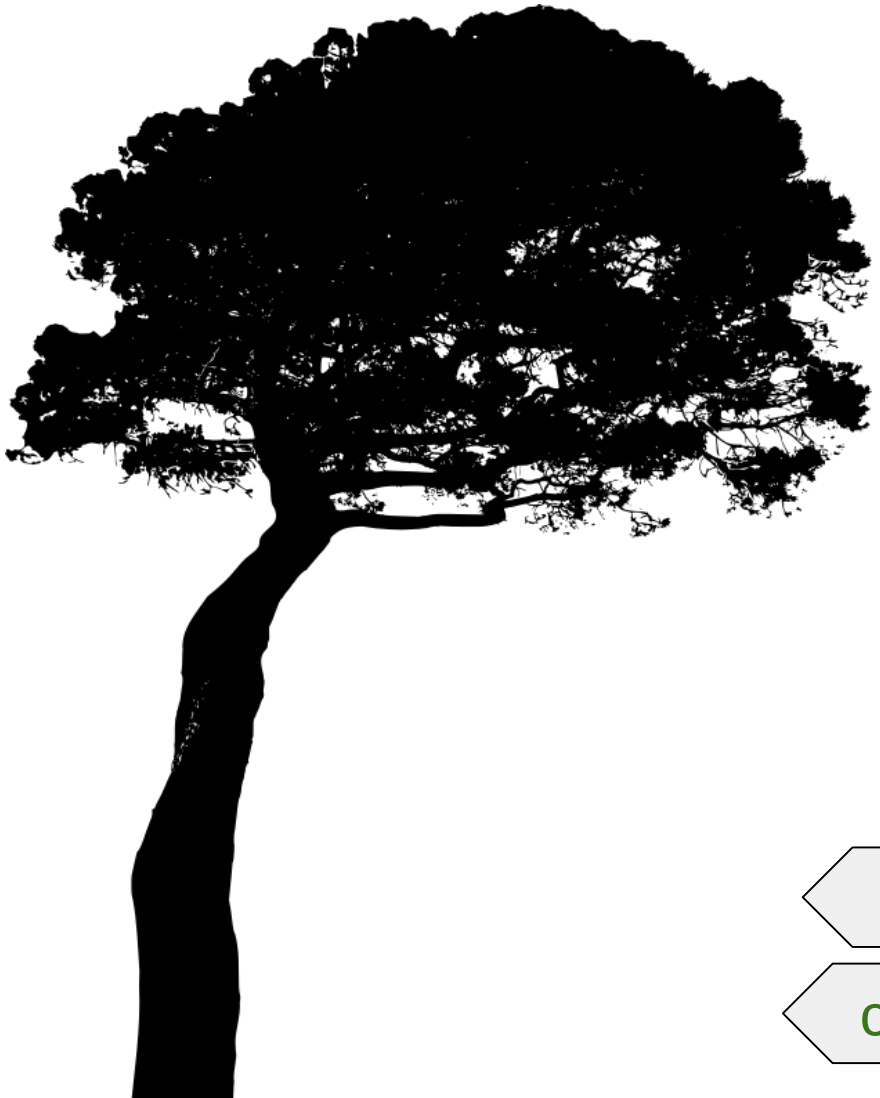
(link)



?

EOL TraitBank

An open digital repository for organism traits



- plant height: 30 m
- leaf area: 1.38 cm^2
- leaf shape: acicular
- plant growth habit: tree
- wood density: 0.5 g/cm^3
- life cycle habit: perennial
- dispersal vector: autochory
- habitat: mediterranean woodland
- conservation status: least concern

Data sources

Databases

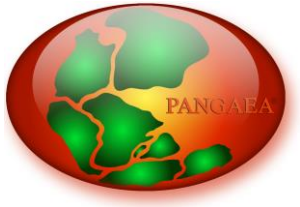
Literature

Natural history collections

Citizen science

Text mining

Legacy/unpublished data



Partners

Evangelos Pafilis

Environments-EOL

This species breeds in mature deciduous or mixed forest and plantations on low hills and mountains, up to 100 m. In central Japan, wooded valleys at lower elevations are preferred. On the wintering grounds, records are from various forest habitats up to 700 m, including mangroves. On migration, it is also recorded from open woodland, suburban parks and gardens in lowlands.

Terpsiphone atrocaudata
Japanese Paradise Flycatcher

habitat ▿ mangrove biome

Environments - …

Data about this record

source	http://eol.org/pages/1051290/details#habitat
measurement method	text mining
contributor	Environments-EOL
measurement remarks	source text: "mangroves"
scientific name	Terpsiphone atrocaudata (Eyton, 1839)
Link to this record	http://eol.org/pages/1051290/data#data_point_11867522

Partners

Evangelos Pafilis

Bioinformatics

ENVIRONMENTS and EOL: identification of Environment Ontology terms in text and the annotation of the Encyclopedia of Life

Evangelos Pafilis[1,*], Sune P. Frankild[2], Julia Schnetzer[3,4], Lucia Fanini[1], Sarah Faulwetter[1], Christina Pavloudi[1], Katerina Vasileiadou[1], Patrick Leary[5], Jennifer Hammock[6], Katja Schulz[6], Cynthia Sims Parr[6,†], Christos Arvanitidis[1] and Lars Juhl Jensen[2,*]

Received November 13, 2014.
Revision received December 23, 2014.
Accepted January 18, 2015.

>500,000 habitat terms for >130,000 taxa

Partners

Anne Thessen

PLOS ONE

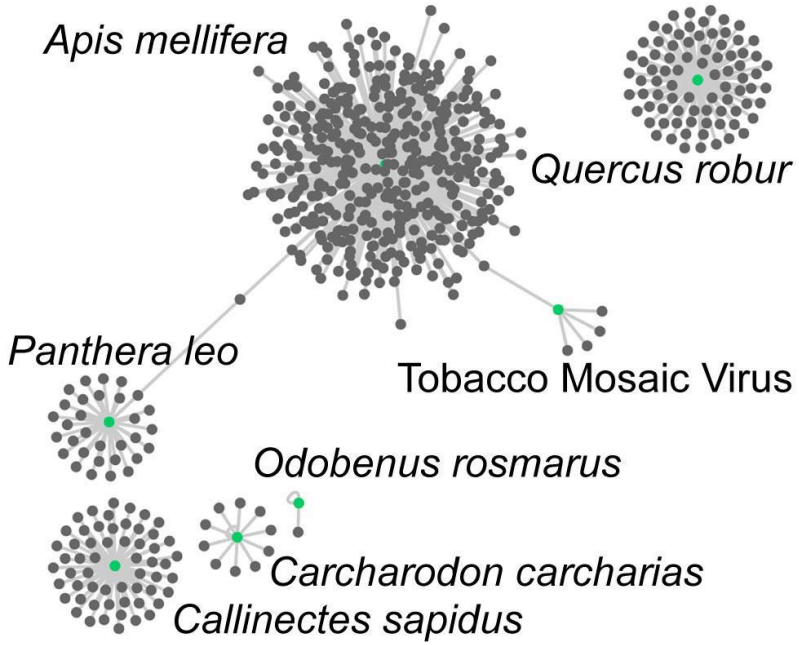
OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Knowledge Extraction and Semantic Annotation of Text from the Encyclopedia of Life

Anne E. Thessen, Cynthia Sims Parr

Published: March 3, 2014 • DOI: 10.1371/journal.pone.0089550



180,000 interaction for 35,000 taxa

Partners

Jorrit Poelen

globalbioticinteractions.org

ELSEVIER

Ecological Informatics

Volume 24, November 2014, Pages 148–159

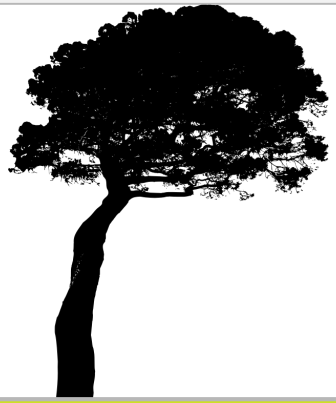
Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets

Jorrit H. Poelen[a], James D. Simons[b], Chris J. Mungall[c]

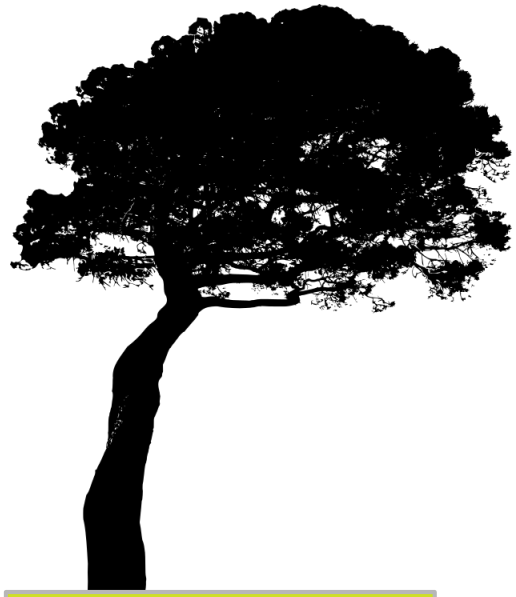
doi:10.1016/j.ecoinf.2014.08.005

>700,000 interaction for >50,000 taxa

Data types（数据类型）

measurements 量度（长、宽、高）	statistics 统计值（平均、最大最小）	facts 特征性状
		
wood density	**average** wingspan	dispersal vector
0.5 g/cm^3	1.58 m	anemochory

Links to ontologies（本体）



本体：一种“形式化的，对于共享概念体系的明确而又详细的说明”

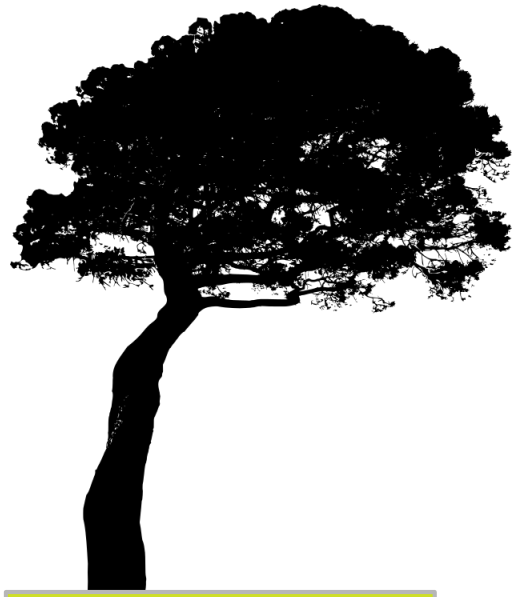
shedability

evergreen

PATO:0001729 shedability
An organismal quality inhering in a bearer by virtue of the bearer's disposition to lose an entitity by natural process.

PATO:0001733 evergreen
A quality inhering in a plant by virtue of the bearer's disposition to retain foliage.

Links to ontologies



shedability

evergreen

The Open Biological and Biomedical Ontologies

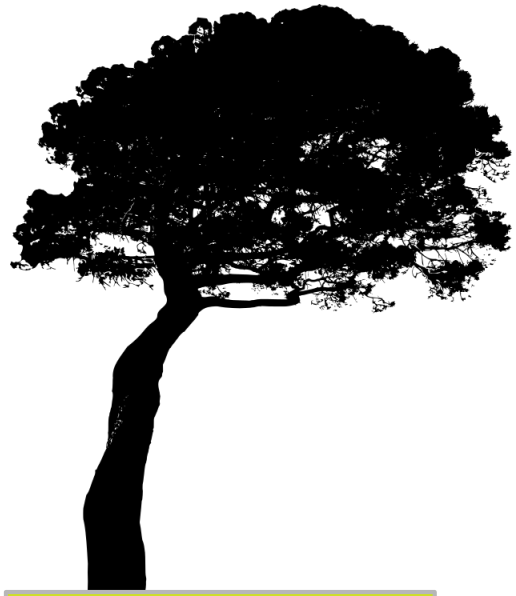
WIKIDATA

PATO:0001729 shedability
An organismal quality inhering in a bearer by virtue of the bearer's disposition to lose an entitity by natural process.

PATO:0001733 evergreen
A quality inhering in a plant by virtue of the bearer's disposition to retain foliage.

Name reconciliation

（关联同一物种不同名字）



shedability

evergreen

EOL:999491 ***Pinus pinea***

a.k.a. *Pinea esculenta*
Pinus maderiensis
Pinus fastuosa

PATO:0001729 **shedability**
An organismal quality inhering in a bearer by virtue of the bearer's disposition to lose an entity by natural process.

PATO:0001733 **evergreen**
A quality inhering in a plant by virtue of the bearer's disposition to retain foliage.

Data integration（数据整合）

Pinus pinea
Pinea esculenta
Pinus maderiensis
Pinus fastuosa

PATO:0001729

shedability
leaf phenology
leaf retention

PATO:0001733

evergreen
non-deciduous
year-round

Data integration

PATO:0001729

PATO:0001733

1 TraitBank Record

Current status

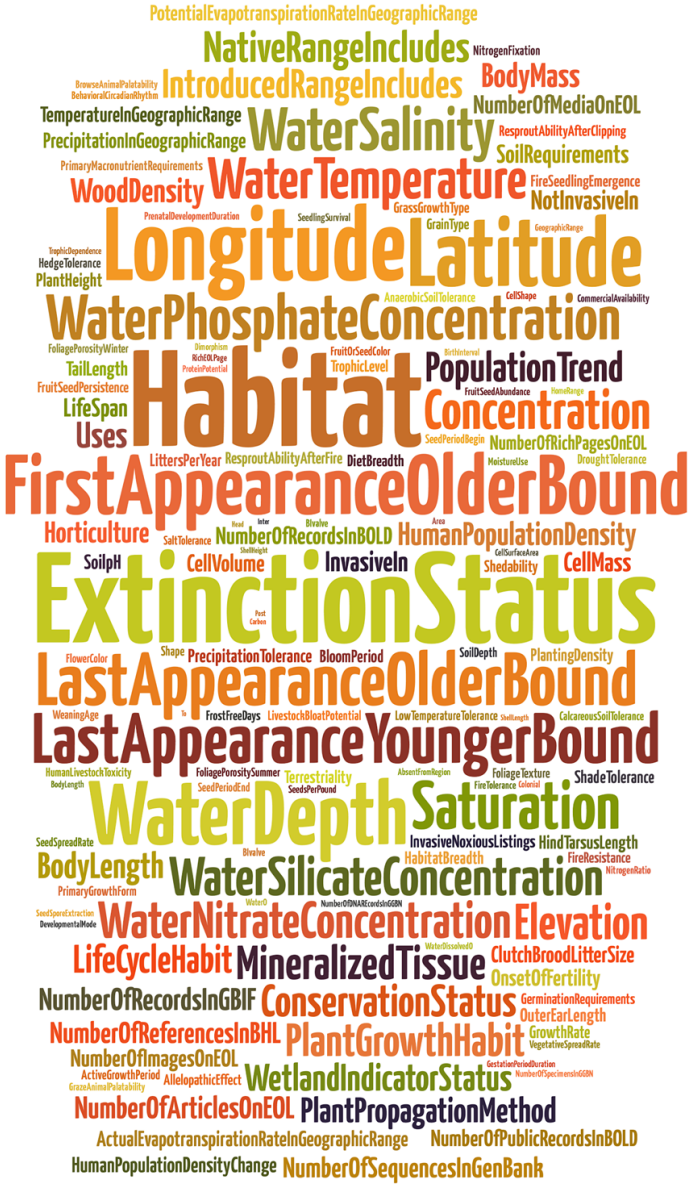
11 M data records

1.7 M taxa with data

340 traits

53 data sets

eol.org/statistics/data



TraitBank is integrated into EOL

Prunus africana

Red Stinkwood learn more about names for this taxon

Overview | Detail | Data | 5 Media | 4 Maps | Names | Community | Resources | Literature | Updates | Worklist

Prunus africana TRUSTED

(cc) BY-NC © Mark Hyde, Bart Wursten and Petra Ballings

Source: Flora of Zimbabwe

see all media
see all maps

EOL has data for 8 traits see all

wood density	0.58 g/cm^3 0.6 g/cm^3 0.62 g/cm^3 more
plant growth habit	tree
habitat	forest mountain terrestrial biome more
elevation	965 m (measurement)
geographic distribution	Africa & Madagascar - Angola Africa & Madagascar - Cameroon Africa & Madagascar - Congo (Brazzaville) more
conservation status	vulnerable
extinction status	extant

TraitBank is integrated into EOL

Prunus africana
Red Stinkwood

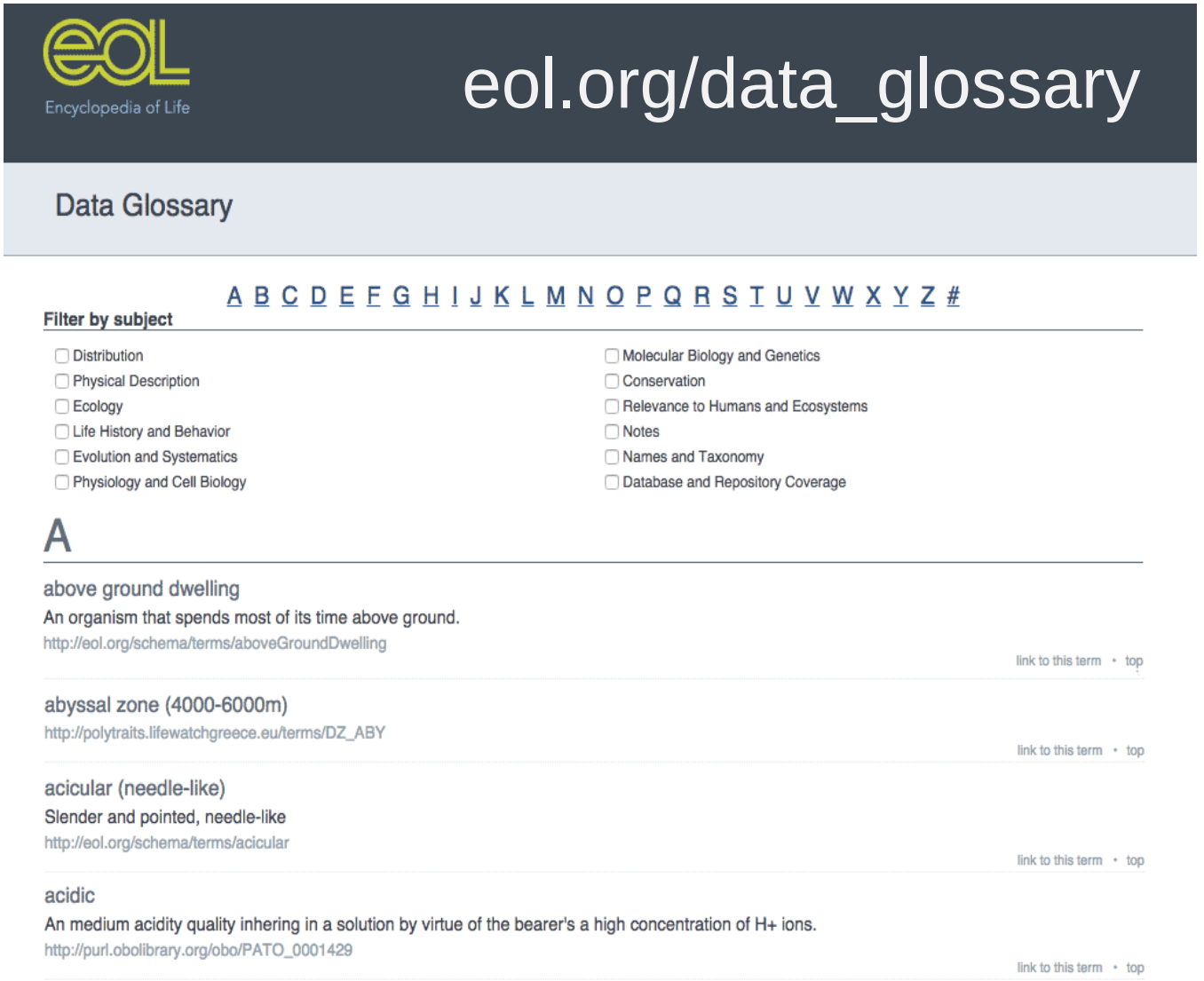
wood density ▼ 0.94 g/cm^3 Global Wood D…

Data about this record

source	Zanne AE, Lopez-Gonzalez G, Coomes DA, Ilic J, Jansen S, Lewis SL, Miller RB, Swenson NG, Wiemann MC, Chave J (2009) Data from: Towards a worldwide wood economics spectrum. Dryad Digital Repository. doi:10.5061/dryad.234
citation	Chave J, Coomes D, Jansen S, Lewis SL, Swenson NG, Zanne AE (2009) Towards a worldwide wood economics spectrum. Ecology Letters 12: 351-366. doi:10.1111/j.1461-0248.2009.01285.x
measurement method	oven dry mass/fresh volume
locality	Africa (extratropical)
Scientific name	Prunus africana
Reference	Goldsmith, B. and D.T. Carter. 1981. The indigenous timbers of Zimbabwe. The Zimbabwe Bulletin of Forestry Research No. 9:x, 406 pp.
Link to this record	http://eol.org/pages/301081/data#data_point_1472441

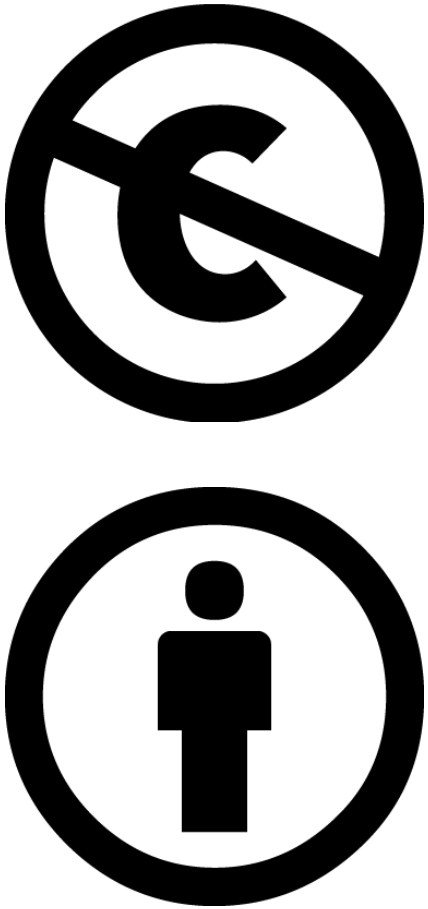
Supplier: Global Wood Density Database

Data glossary（数据术语表）



Open access data（数据开放访问）

- Databases
- Literature
- Natural history collections
- Citizen science
- Text mining



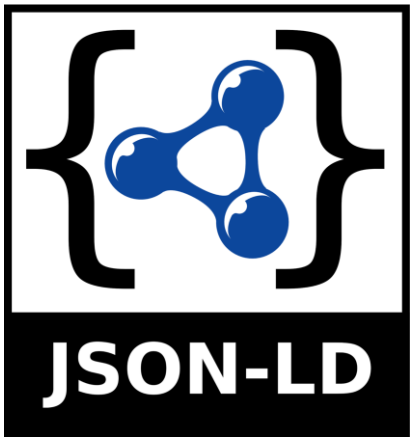
Search & download

Data Search Service

eol.org/data_search/taxon_id

JSON-LD

eol.org/api/traits



Current data priorities
当前数据的优先级

Measures of body size

How big/heavy is it?

Distribution & habitat

Where does it live?

Trophic ecology

What does it eat? What eats it?

GEO BON Essential Biodiversity Variables

Phenology, demographic & physiological traits, migratory behavior, natal dispersal distance (Pereira et al. 2013)

Applications

Ecosystem modeling

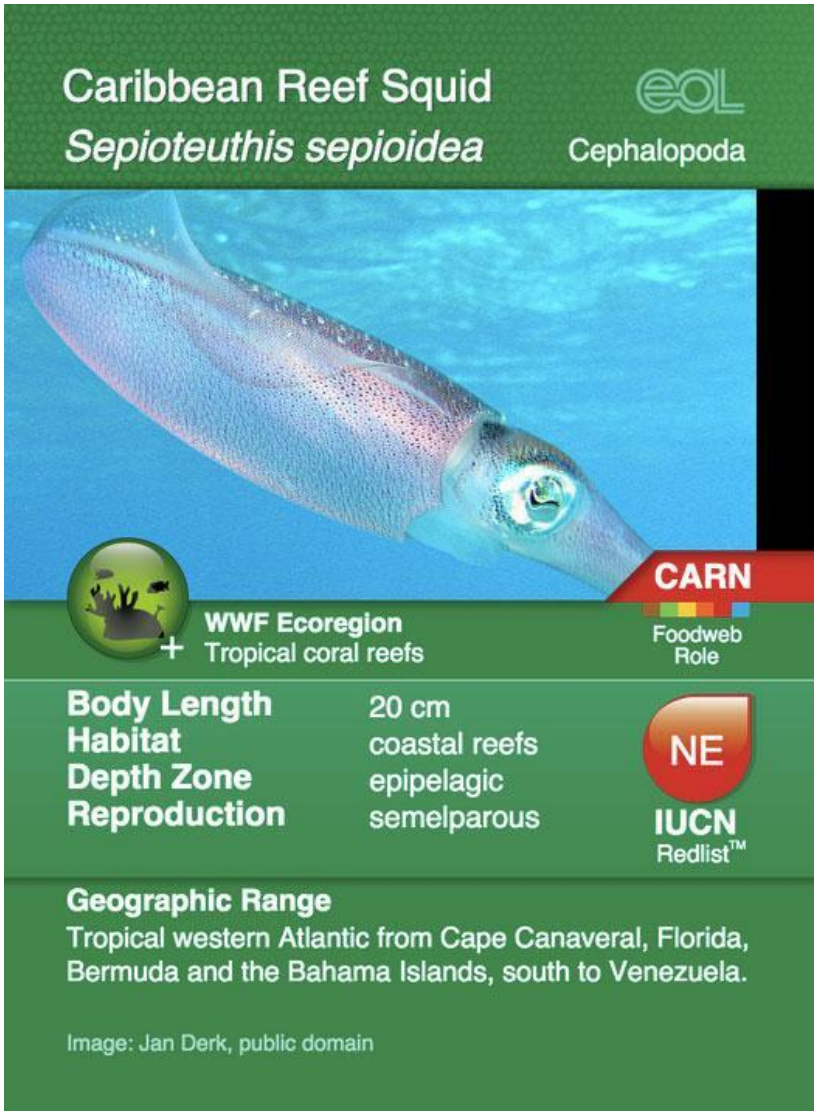
Phylogenetics/character evolution

Google Knowledge Graph

Data in the classroom

Biodiversity Apps/Games

Applications



Koala

Animal

The koala is an arboreal herbivorous marsupial native to Australia. It is the only extant representative of the family Phascolarctidae, and its closest living relatives are the wombats. Wikipedia

Scientific name: Phascolarctos cinereus

Trophic level: Herbivorous

Lifespan: 13 – 18 years (In Wild)

Mass: 8.8 – 33 lbs (Adult)

Higher classification: Phascolarctos

Gestation period: 30 – 35 days

Body length: 2 – 2.8 ft. (Adult)

Sources include: EOL

Feedback

Google

Koala

Animal

The koala is ar
the only extant
closest living re

Scientific nam

Trophic level:

Lifespan: 13 –

Mass: 8.8 – 33

Higher classif

Gestation period: 30 – 35 days

Body length: 2 – 2.8 ft. (Adult)

Sources include: EOL

Feedback

Higher classification:

Gestation period: 30 –

Body length: 2 – 2.8 ft

Sources include: EOL

TraitBank Summary

- Serves data across tree of life
- Complements specialized repositories, fill gaps
- Mobilizes legacy data, poorly structured data, citizen science data (data from diverse sources)
- Experiments with text-mining
- Discovery tool for specialized repositories
- Documents poorly known taxa
- Serves structured data in standardized format for a variety of applications, e.g., identification tools

概述：基于文本的数据挖掘，将非结构化的数据（文字描述类型）结构化后，服务于各种应用程序，如鉴定工具等

EOL TraitBank

An open digital repository for organism traits

Bob Corrigan

Director of Operations, Encyclopedia of Life
Smithsonian Institution
National Museum of Natural History

4th Conference on Biodiversity Informatics in China
November 4, 2015