Discover Life Bee Checklist Archive

Poelen, JH https://orcid.org/0000-0003-3138-4118
Seltmann, KC https://orcid.org/0000-0001-5354-6048

2023-08-29

Abstract

Digital biodiversity knowledge resources are increasingly available openly on the internet. Some of these potentially valuable resources are still actively curated, whereas others may have lost their maintenance/curators due to life events, funding, or a change in institutional policy. This data publication records a snapshot of the authoritative resource on the biodiversity of bees: Ascher, J. S. and J. Pickering. 2022. Discover Life bee species guide and world checklist (Hymenoptera: Apoidea: Anthophila). http://www.discoverlife.org/mp/20q?guide=Apoidea_species Draft-55, 17 November 2020. The reason for making this snapshot is to provide a citable data package containing the Discover Life Bee Checklist for use in data synthesis and integration workflows. This data package is versioned and made verifiable using Preston, a biodiversity data tracker. With this publication, verifiable versions of the Discover Life Bee Checklist can now be cited and copied regardless of physical location.

Contents

:warning: work in progress

Introduction

Life on Earth is supported by a complex and diverse network of interactions between organisms and their surroundings. With the advancement of digital storage, processing, and networking technologies, (community) scientists now have the ability to access digital datasets that document various aspects of life on Earth through the internet. However, there is growing evidence indicating that these easily accessible digital datasets might eventually become unavailable due to broken links or undergo changes over time, a phenomenon known as "linkrot" or "content drift" (Elliott, Poelen, and Fortes 2020, 2023). In order to mitigate the risk of losing or altering valuable digital biodiversity datasets, researchers are employing content-based data tracking methods. Here, these methods are applied to a widely used digital resource for bee names, specifically the DiscoverLife Bee Checklist (Ascher and Pickering 2022).

The Discover Life Bee Checklist is the most comprehensive checklist for bees in the West and is commonly referenced for ecological research. It is constructed via a collaboration between John S. Ascher and John Pickering, drawing on taxonomic publications and prior work by many people. The list is periodically peer-reviewed by ITIS as part of the GBIF-supported World Bee Checklist project. For more information about the checklist and its sources, see https://www.discoverlife.org/mp/20q?act=x_guide_credit&guide=Apoidea_species.

Methods

To help version a snapshot of the Discover Life Bee Checklist, the following openly available tools were used: bash, Preston, grep, xmllint, cut, and xargs. With these tools the following archiving workflow was implemented:

```
#!/bin/bash
#
# Makes an archive of DiscoverLife Bee checklist and associated species pages.
#

preston track "https://www.discoverlife.org/mp/20q/?act=x_checklist&guide=Apoidea_species&flags=HAS" \
 | grep hasVersion\
 | preston cat\
 | xmllint --html --xpath '//table//tr/td/i/a/@href' -\
 | cut --delimiter '"' -f2\
 | sed 's+^+https://www.discoverlife.org+g'\
 | xargs -L100 preston track

# retry previously failed web requests, if needed.
```

```
15 preston ls -l tsv\
16  | grep well-known\
17  | grep hasVersion\
18  | cut -f1\
19  | xargs preston track
```

In this workflow, on line 6, the command `preston track` captures a snapshot of HTML pages that contain references to pages for various bee species. The result of this tracking process is a detailed stream of statements describing the tracking steps. This output is then passed to grep hasVersion through a Linux pipe, which filters and selects only the statements that connect the web addresses with the discovered content.

Subsequently, the content associated with these statements is streamed to the standard output using the `preston cat` command. From this streamed content, URLs to the species pages are generated. This is done by first extracting relevant HTML fragments using an XPath query. Then, these fragments are transformed into URLs using a combination of string parsing (using the cut –delimiter '"' -f2 command) and stream editing (with `sed 's+^+https://www.discoverlife.org+g'`).

The resulting URLs, which lead to pages about bee species, are organized into blocks of 100 URLs each and tracked using the Preston tool. The workflow includes a retry procedure to account for potential failures in making web requests. This ensures that web locations that initially fail to provide content are retried to compensate for any issues.

Results

The resulting archive can access a versioned copy of the Discover Life Bee Checklist. The archive contains over 20k HTML pages that appear to be consistently structured. This consistent structure allows scripts or other computer programs to transform the data into a format suitable for reuse automatically.

Table 1: First three DiscoverLife Bee Checklist HTML resources tracked. The first contains the index page of species pages. The following two are locations, and associated content identifiers, to species pages associated with *Andrena angustior* and *Andrena angusticrus*. This table was generated using `preston alias -l tsv | tail -n3 | tac | cut -f1-3 | mlr --hi --itsvlite --omd cat` .

discoverlife url	content id
…guide=Apoidea_species&flags=HAS	sha256:c4f…
…Andrena+angustior	sha256;3091…
…Andrena+angusticrus	sha256:afe0…

The current content identifiers of this versioned package of DiscoverLife Bee Checklist html resources are:

hash://sha256/86e7ce5f3df9a136a2957de5655261c007b95e217b2f0901988ffb39ee0230fe

hash://md5/55fe2b12ab306704ce332d97723b95af

Example 1. List Most Frequently Appearing Bee Subgenus Names

DiscoverLife species pages document subgenera associated with bee species in html fragments such as:

```
<small>Subgenus: <a
 ↪  href="/mp/20p?see=Archianthidium&amp;name=Trachusa&amp;
 ↪  flags=subgenus:"><i>Archianthidium</i></a></small>
```

The html fragment above was seen at a page describing *Trachusa forcipata* with content id hash://sha256/ce144a314ef4bafa714f69215065447309109 35a870786964506dc18c65349dd.

To query for the top 10 most frequently appearing subgenera appearing in the pages, you can use:

```
preston ls\
 --remote https://linker.bio,https://softwareheritage.org\
 --anchor
  ↪  hash://sha256/86e7ce5f3df9a136a2957de5655261c007b95e217b2f0901988ffb39ee0230fe
  ↪  \
 -l tsv\
 | grep -v well-known\
 | grep hasVersion\
 | cut -f3\
 | preston cat\
 | grep "Subgenus:"\
 | sed 's+<br>.*<i>++g'\
 | sed 's+</i></a></small>++g'\
 | sort\
 | uniq -c\
 | sort -nr\
 | head
```

The result is shown in the table below.

Table 2: Top 10 most frequent appearances of (likely) subgenus names in the bee species pages ordered by decreasing frequency:

frequency	subgenus
5600	None
765	Uncertain

frequency	subgenus
448	Perdita
403	Dialictus
259	Hemihalictus
209	Eutricharaea
179	Ctenonomia
161	Homalictus
152	Anthidium
151	Lasioglossum

Example 2. List Bee Hosts

The DiscoverLife Bee checklist contains information about (plant) hosts associated with specific bees. This information is captured in html snippets such as:

```
<p><table width="80%"><tr><td><a name="Hosts"><table
 ↪  cellspacing="0" cellpadding="0" border="0"><tr><td
 ↪  colspan=2"><b>Hosts</b> &middot; <a
 ↪  href="/mp/20m?kind=Agapostemon+texanus&m_i=h&m_order=0">map
 ↪  </a></td></tr><tr><td><u>Family</u></td><td><u>Scientific
 ↪  name</u> <font size="-1" face="sans-serif">@ source
 ↪  (<u>records</u>)</font></td></tr><tr><td valign="top"><a
 ↪  href="/20/q?search=Asteraceae">Asteraceae</a>  </
 ↪  td><td valign="top" nowrap><a
 ↪  href="/20/q?search=Achillea+millefolium">Achillea
 ↪  millefolium</a><font size="-1" face="sans-serif"> @ UCMS_ENT
 ↪  <a href="/mp/20l?id=UCMS_ENT00058904;UCMS_ENT00058903">(2)</
 ↪  a></font></td></tr>
```

as extracted from line 538 of content associated with DiscoverLife Bee page on *Agapostemon texanus*.

With this, the script below can be constructed to extract hosts from this particular species page:

```
1 preston cat
   ↪  'hash://sha256/7168d15fe822bc6770954b9e3a3b64b62f05ccad636c293e9d5a07d6fb173ddc'
   ↪  \
2  | xmllint\
3  --html\
4  --xpath "//a[@name='Hosts']/following-sibling::*//td/a/text()"\
5  -\
6  | grep -oE "[A-Z][a-z]{1,}[ ].*"
```

where, `preston cat ...` streams a species page with content id `hash://sha256/7168d...` and selects associated host species by combining an XPath query (line 3) with

a regular expression (line 4).

This script was used to generate the following list of known hosts of *Agapostemon texanus*, as claimed by (Ascher and Pickering 2022):

```
Achillea millefolium
Aletris farinosa
Arnica sp
Aster simplex
Aster sp
Astragalus racemosus
Baccharis salicina
Baileya multiradiata
Barbarea vulgaris
Beta vulgaris
Bidens ferulifolia
Blephilia ciliata
Chrysanthemum leucanthemum
Chrysothamnus sp
Chrysothamnus viscidiflorus
Cichorium intybus
Cirsium sp
Cirsium vulgare
Cleome serrulata
Cleome sp
Convolvulus sepium
Conyza canadensis
Coreopsis sp
Ericameria nauseosa
Erigeron annuus
Erigeron leiomerus
Eriogonum sp
Erysimum repandum
Eupatorium purpureum
Flaveria campestris
Fragaria virginiana
Glaucium flavum
Grindelia sp
Grindelia squarrosa
Helianthus annuus
Helianthus anomalus
Helianthus sp
Heterotheca inuloides
Heterotheca subaxillaris
Hieracium sp
Horkelia sp
Kalmia latifolia
```

```
Larrea tridentata
Lathyrus japonicus
Leucanthemum vulgare
Limonium carolinianum
Machaeranthera bigelovii
Machaeranthera sp
Madia elegans
Malus pumila
Medicago sativa
Petrophyton caespitosum
Phacelia sp
Plantago lanceolata
Poinsettia heterophylla
Prosopis glandulosa
Prosopis sp
Raphanus raphanistrum
Ratibida columnifera
Rosa rugosa
Rubus sp
Rubus spp
Salvia carduacea
Sclerocactus wrightiae
Senecio sp
Solidago tenuifolia
Sphaeralcea sp
Taraxacum campylodes
Tephrosia virginiana
Teucrium canadense
Trifolium hybridum
Trifolium repens
Verbena sp
Vernonia noveboracensis
```

The examples above show two applications of data extraction from (Ascher and Pickering 2022): extracting most frequently appearing subgenera names, and extracting host plants for a specific species page.

Discussion

Biodiversity datasets are available online as html pages, or structured in other digital formats. In this publication, one such resource (Ascher and Pickering 2022) was tracked and packaged into a citable biodiversity dataset containing over 20k HTML resources. The data tracking method may be applied to other currently available network-accessible biodiversity datasets in an effort to turn webpages into versioned digital research objects.

References

Ascher, John S., and John Pickering. 2022. "Discover Life Bee Species Guide and World Checklist (Hymenoptera: Apoidea: Anthophila)." DiscoverLife. http://www.discoverlife.org/mp/20q?guide=Apoidea_species.

Elliott, Michael J., Jorrit H. Poelen, and José A. B. Fortes. 2020. "Toward Reliable Biodiversity Dataset References." *Ecological Informatics* 59 (September): 101132. https://doi.org/10.1016/j.ecoinf.2020.101132.

———. 2023. "Signing Data Citations Enables Data Verification and Citation Persistence." *Scientific Data* 10 (1). https://doi.org/10.1038/s41597-023-02230-y.