DiscoverLife Bee Checklist Archive

Poelen, JH

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 an authoritive 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 The reason for making this snapshot is to provide a citable data package containing the DiscoverLife 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 DiscoverLife Bee Checklist can now be cited and copied regardless of their physical location.

Contents

:warning: work in progress

Introduction

Life on earth is sustained through a complex, and diverse, web of relationships between organisms and their environment. Now that digital storage, processing and networking technologies are within reach of (community) scientists, digital datasets documenting life on earth are increasingly available through the internet. However, evidence suggestions these network accessible digital datasets are likely to become unavailable due to linkrot, or change due to content drift [@elliott2020, @elliott2023]. To help reduce the risk of dataloss (or change) of valuable digital biodiversity datasets, content-based data tracking methods are applied to a commonly used digital biodiversity resource, the DiscoverLife Bee Checklist [@ascher2022].

Methods

To help version a snapshot of the DiscoverLife 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:

```
 1 #!/bin/bash
 2 #
 3 # Makes an archive of DiscoverLife Bee checklist and associated species pages.
 4 #
 5 
 6 preston track "https://www.discoverlife.org/mp/20q/?act=x_checklist&guide=Apoidea_species&fl
 7  | grep hasVersion\
 8  | preston cat\
 9  | xmllint --html --xpath '//table//tr/td/i/a/@href' -\
10  | cut --delimiter '"' -f2\
11  | sed 's+^+https://www.discoverlife.org+g'\
12  | xargs -L100 preston track
13 
14 # re-try previously failed web requests
15 preston ls -l tsv\
16  | grep well-known\
17  | grep hasVersion\
18  | cut -f1\
19  | xargs preston track
```

In this workflow, `preston track` in line 6 take a snapshot of an html pages that contains references to all bee species pages. The output of this tracking process is a stream of statement describing the tracking process in great detail. This output is fed into `grep hasVersion` using a linux pipe to selects only statements that associate the web location with the content that was found. Following, the associated content is streamed to stdout (or standard output) using `preston cat`. Following, URLs to species pages are generated from this streamed content by extracting a relevant html fragments using a xpath query. Then, this fragment is transformed into a URLs using string parsing (i.e., `cut --delimiter '"' -f2\`) and stream editing (i.e., `sed 's+^+https://www.discoverlife.org+g'`). The resulting URLs of the bees species pages are then tracked, in blocks on 100 URLs, by Preston. To help compensate for likely web request failures, the workflow was completed with a retry procedure for web locations that failed to successfully provide content initially.

Results

The resulting archive can be used to access a versioned copy of discover life. The archive contains over 20k HTML pages that appear to be consistently structured. This consistent structure allow for scripts, or other computer programs, to automatically transform the data into a format suitable for reuse.

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:">
```

The html fragment above was seen at a page describing *Trachusa forcipata* with content id hash://sha256/ce144a314ef4bafa714f6921506544730910935a870786964506dc18c65349dd.

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

```
preston ls\
 --remote https://linker.bio,https://github.com/Big-Bee-Network/discoverlife-bee-archive/raw
 --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
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=
```

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:

```
preston cat 'hash://sha256/7168d15fe822bc6770954b9e3a3b64b62f05ccad636c293e9d5a07d6fb173ddc
 | xmllint\
 --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 taxanus, as claimed by [@ascher2022]:

```
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 [@ascher2022]: 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 [@ascher2022] 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.