Getting data

2020-11-06

BIODIVERSITY.AQ

Biodiversity data from the field to research

Maxime Sweetlove

SCAR Antarctic Biodiversity Portal

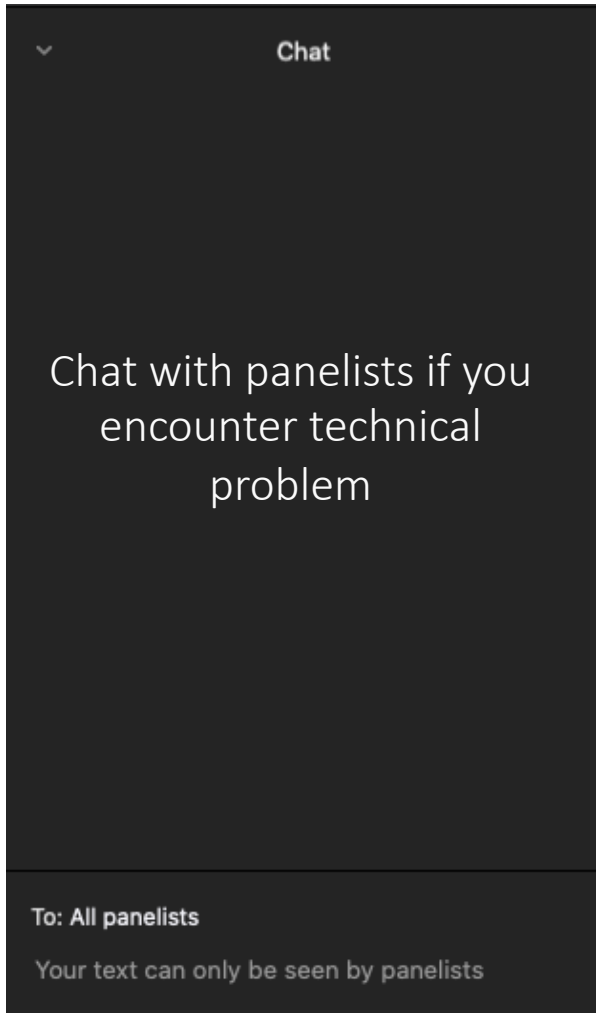
We are recording this seminar

Let us know if you prefer not to be recorded.

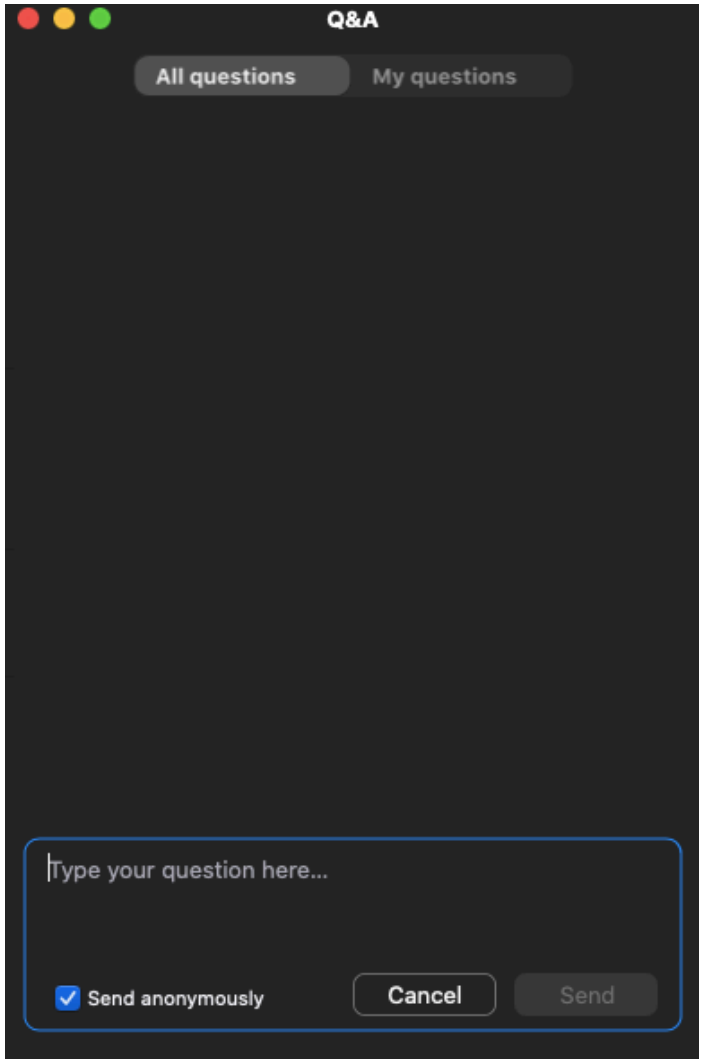
Code of Conduct

- Be respectful
- We will follow the principles of the rOpenSci Code of Conduct
  - https://ropensci.org/code-of-conduct/

Using Zoom in this webinar

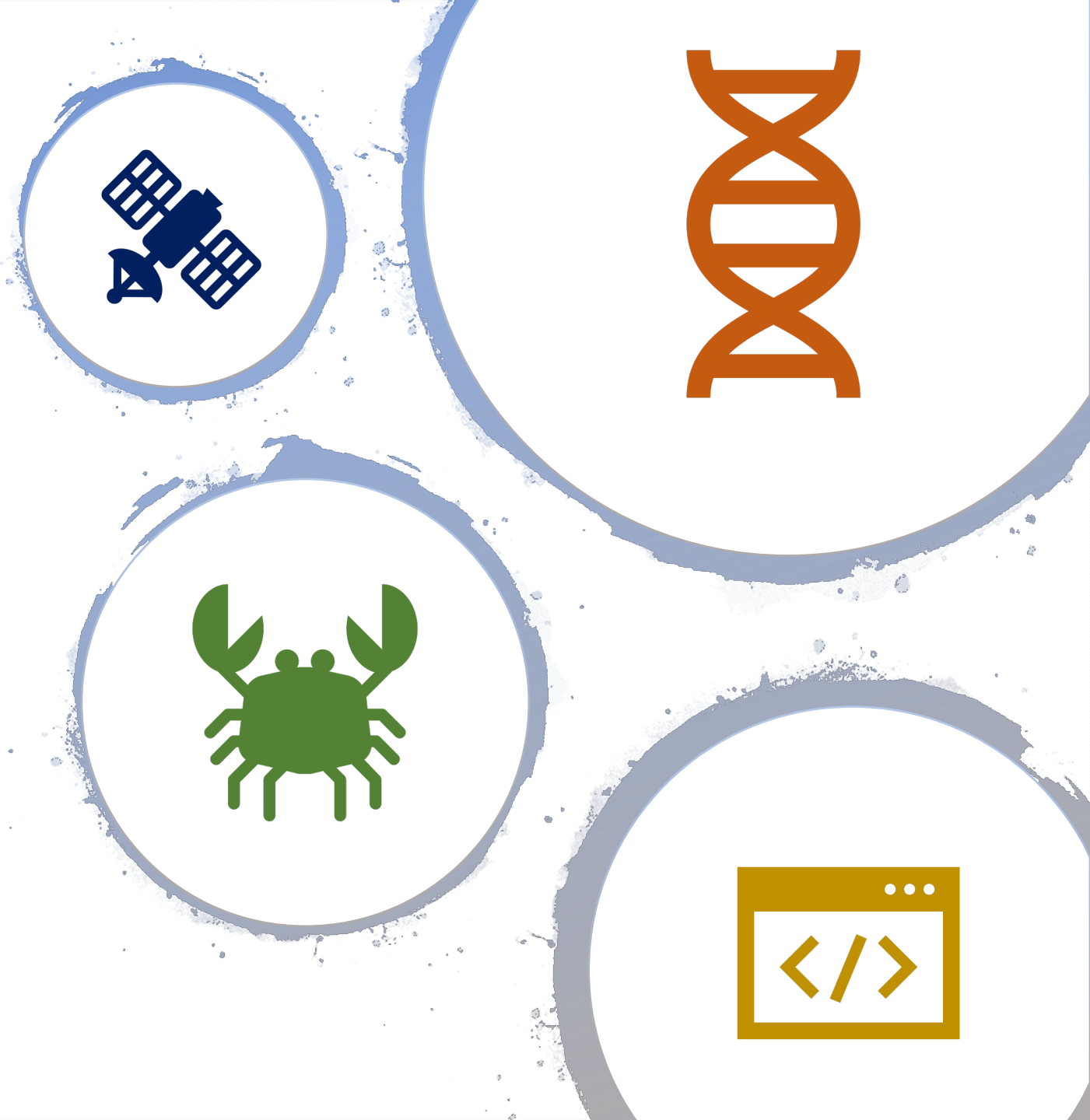


Ask questions using Q&A feature
or
raise your hand



In this webinar:

- Focus on retrieving data for ecological or biodiversity research
  - Species occurrences
  - Genetic data
  - Environmental data
- Advanced ways for big data
  - web services
  - Demo: R for occurrences
  - Demo: R for sequence data
- Many examples on where to get data
  - The slides will be put online later, you can retrieve all the links from there!



Good reasons to look for more data

- Increase sample size
- Broaden geographic scope
  - Sample more locations
- Broaden temporal scope
  - Compare with the past
- find explanatory variables
  - Correlations with environment or climate
- Meta-analysis
  - Patterns across studies
  - Estimate effect size

Article | Published: 18 March 2020

Tracking of marine predators to protect Southern Ocean ecosystems

Mark A. Hindell ✉, Ryan R. Reisinger, [...] Ben Raymond

Nature **580**, 87–92(2020) | Cite this article

Abstract

Southern Ocean ecosystems are under pressure from resource exploitation and climate change[1,2]. Mitigation requires the identification and protection of Areas of Ecological Significance (AESs), which have so far not been determined at the ocean-basin scale. Here, using assemblage-level tracking of marine predators, we identify AESs for this globally important region and assess current threats and protection levels. Integration of more than 4,000 tracks from 17 bird and mammal species reveals AESs around sub-Antarctic islands in the Atlantic and Indian Oceans and over the Antarctic continental shelf. Fishing pressure is disproportionately concentrated inside AESs, and climate change over the next century is predicted to impose pressure on these areas, particularly around the Antarctic continent. At present, 7.1% of the ocean south of 40°S is under formal protection, including 29% of the total AESs. The establishment and regular revision of networks of protection that encompass AESs are needed to provide long-term mitigation of growing pressures on Southern Ocean ecosystems.

BIOLOGICAL REVIEWS
Cambridge Philosophical Society

Original Article

A meta-analysis of human disturbance impacts on Antarctic wildlife

Bernard W. T. Coetzee ✉, Steven L. Chown

First published: 28 April 2015 | https://doi.org/10.1111/brv.12184 | Citations: 25

ABSTRACT

Evidence-based assessments are increasingly recognized as the best-practice approach to determine appropriate conservation interventions, but such assessments of the impact of human disturbance on wildlife are rare. Human disturbance comprises anthropogenic activities that are typically non-lethal, but may cause short- and/or longer-term stress and fitness responses in wildlife. Expanding human activity in the Antarctic region is of particular concern because it increases the scope and potential for increased human disturbance to wildlife in a region that is often thought of as relatively untouched by anthropogenic influences. Here, we use a meta-analytical approach to synthesise research on human disturbance to wildlife over the last three decades in the Antarctic and sub-Antarctic region. We combine data from 62 studies across 21 species on the

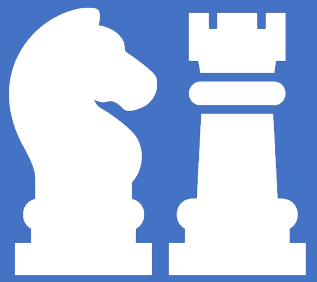
Problem

- Most data repositories are great at archiving data...

  ... but less ideal for discovering data

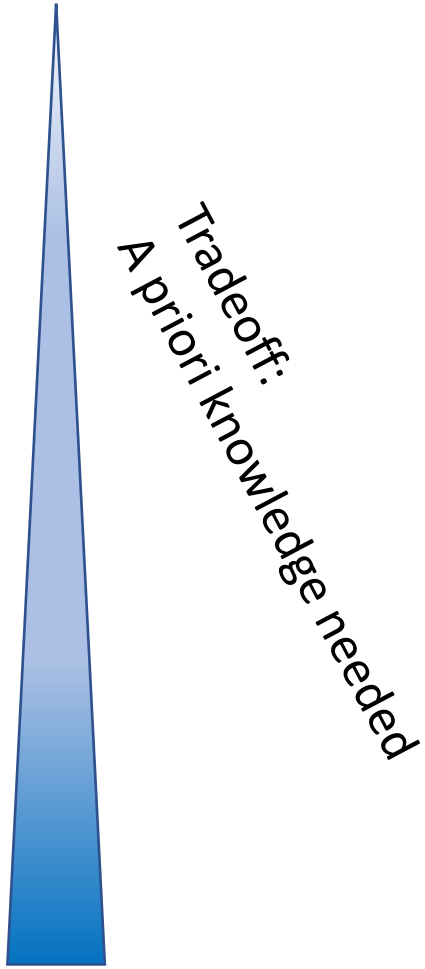
- There are many different data repositories
  - >2600 repositories listed on re3data.org

→ Search strategy

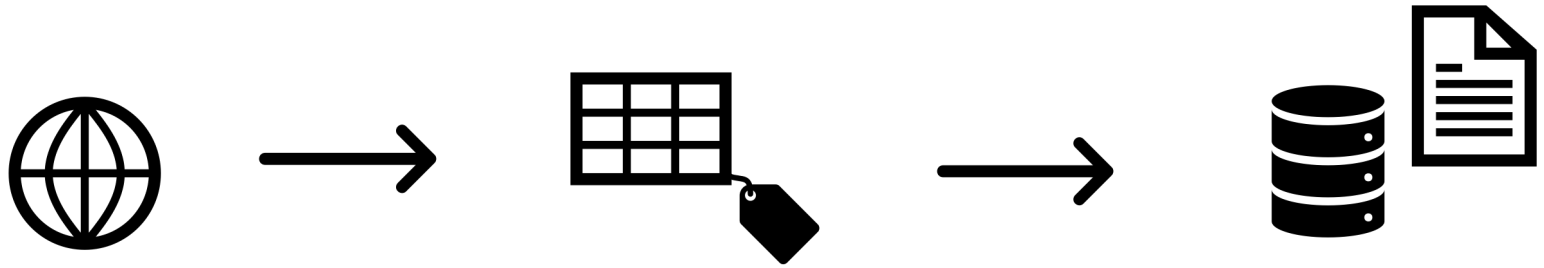


Data search strategies

- Broad-scale
  - Don’t know where to find data
  - The whole internet
  - General purpose data repositories
- Fine-scale
  - You know where/what to look for
  - thematic data repository
  - Specific and standardized data



Broad-scale: internet search engines



- Web Crawlers = browse the web using tags in the website code
- schema.org: tag system for webpages
- Examples:

http://scholar.google.com

Dataset Search http://datasetsearch.research.google.com

Broad-scale: general purpose databases

- Registries for all sorts of data, all kinds of formats
- Usually free text search (=keywords)
- Examples:

zenodo https://zenodo.org

figshare https://figshare.com

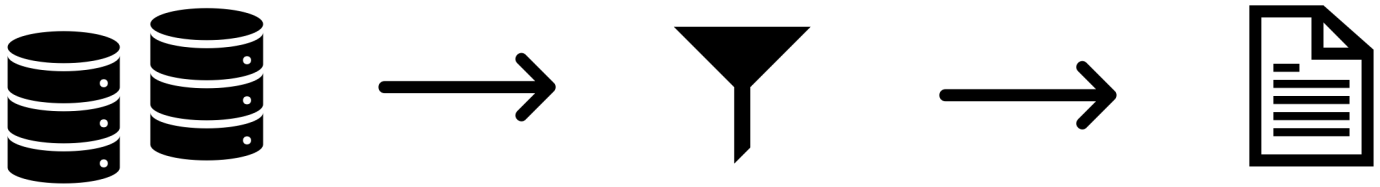
DataONE https://dataone.org

PANGAEA. https://pangea.de

Broad-scale: finding repositories

- Re3data.org
  - Service of DataCite (organization that assigns DOIs)
  - Registry of data repositories

Fine-scale searches



- Look for data where it is stored
  - Repository in known before hand
  - Query the database
- General data repositories
  - e.g. occurrences, DNA sequences,...
- thematic data repositories
  - Antarctic biodiversity, birds, ocean biodiversity,..

Focus for the rest of this webinar

EGABI

- The SCAR Expert Group on Antarctic Biodiversity Informatics
- Great resource
  - Community building
  - training, example workflows, tutorials
  - Software tools
  - Data availability
  - analytical and collaborative platforms

Expert Group on Antarctic Biodiversity Informatics (EGABI)

- established in 2012, supports the Antarctic and Southern Ocean biodiversity science community:
  - encouraging collaboration, disseminating information and advice across the SCAR and broader Antarctic science communities
  - hosting workshops, discussion fora, and other community engagement and capacity building
  - working with other groups in SCAR and elsewhere to develop analytical tools and synthesized data products of value to the community
  - engaging in biodiversity science projects

Contact: Ben Raymond (Chief officer), Anton Van de Putte (Deputy)

Some EGABI initiatives:

Data laundry slack channel

SCAR-EGABI Tools for Southern Ocean Spatial Analysis and Modelling - Course

2-6 September 2019, KULeuven, Leuven, Belgium

Relevant types of data for ecology

Biodiversity occurrences

- Species observations
- Date + place

Nucleotide sequences

- Species markers, genes, genomes
- Metabarcodes, metagenomes

Environmental data

- Geology, climate, chemo-physical
- correlations

Biodiversity occurrences

- Species or taxon identification
  - Usually linked to a taxonomic backbone (=persistent ID to taxonomic name)

    + date

    + location
- Examples:
  - Bird sightings
  - Seal GPS tracks
  - Plankton counts

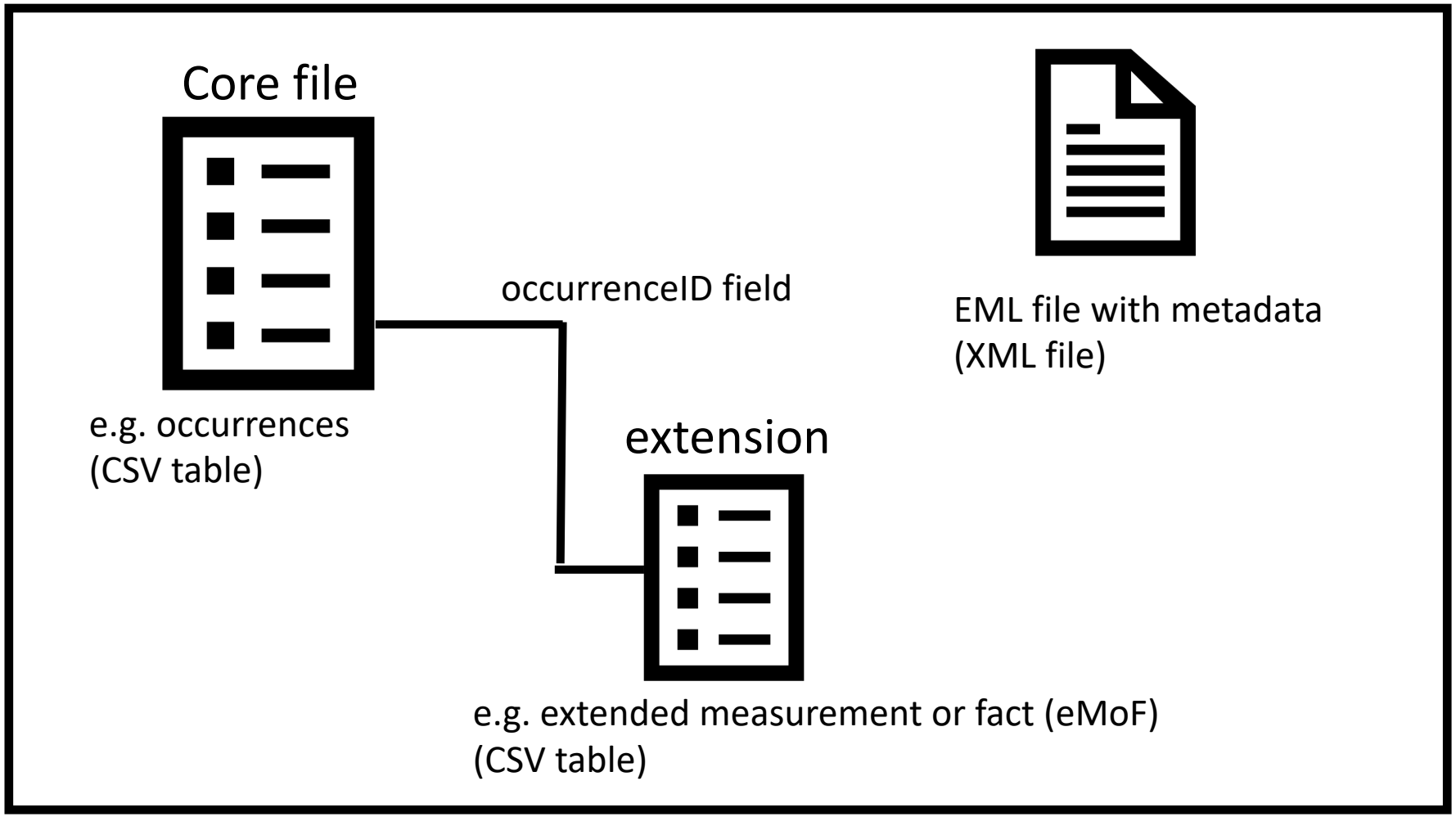
Biodiversity occurrences

- Much open data available
- Endless possibilities:
  - Species distributions
  - Identify specimen in natural history collections (basisOfRecord = “PreservedSpecimen”)
  - Compare to datasets of the past
  - Enrich your own data or do a meta analysis
  - …

Biodiversity occurrences

- Most common data standard: DarwinCore

DarwinCore archive = folder

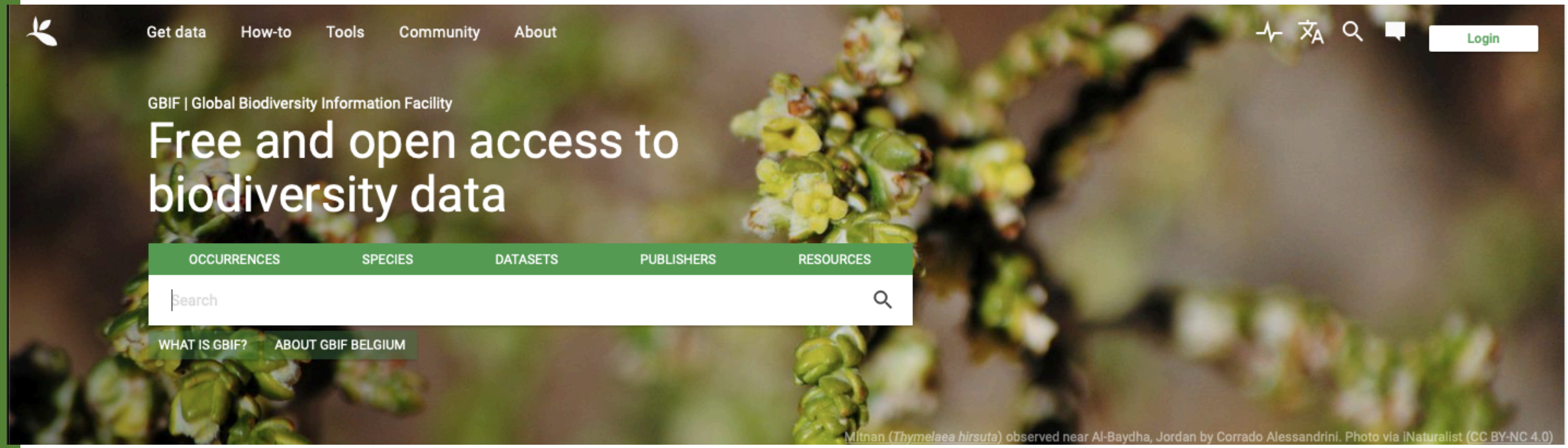


DEMO

Example:
Global Biodiversity Information Facility (GBIF)

Case study:

Cryptopygus antarcticus (an Antarctic springtail):
What is it's known biogeographic range?



https://www.gbif.org/

DEMO

www.gbif.org

Get data How-to Tools Community About Login

GBIF | Global Biodiversity Information Facility

Free and open access to biodiversity data

OCCURRENCES SPECIES DATASETS PUBLISHERS RESOURCES

Search

WHAT IS GBIF? ABOUT GBIF BELGIUM

Goitered gazelle (*Gazella subgutturosa*) observed near Salyan, Azerbaijan by annabutrim. Photo via iNaturalist (CC BY-NC 4.0)

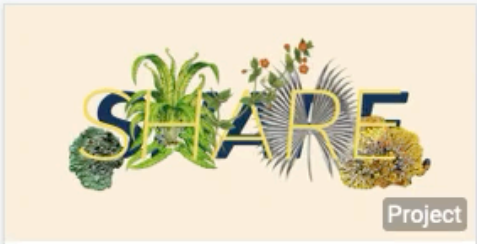
Occurrence records
1,626,808,936

Datasets
54,686

Publishing institutions
1,665

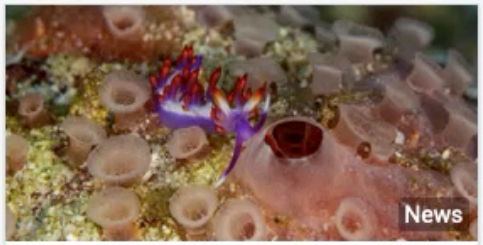
Peer-reviewed papers using data
5,078

New task group to enhance GBIF-enabled research on species linked to human diseases



New video: an invitation to the private sector

BID call for proposals: Caribbean 2020



BID call for proposals: Pacific 2020

DEMO

www.gbif.org/search?q=Cryptopygus%20antarcticus%20

Get data How-to Tools Community About Login

Cryptopygus antarcticus

EVERYTHING OCCURRENCES SPECIES DATASETS PUBLISHERS RESOURCES

Antarctica

1,374,843 occurrences about | 39,405 occurrences published

Cryptopygus antarcticus Willem, 1901

Species

Classification : Animalia › Arthropoda › Entognatha › Collembola › Isotomidae › Cryptopygus

Accepted | Species | 341 occurrences

DATASETS 26 RESULTS

Taxonomy of the *Cryptopygus* complex. III. The revision of South African species of *Cryptopygus* and Isotominella (Collembola, Isotomidae)

Checklist dataset

This dataset contains the digitized treatments in Plazi based on the original journal article Potapov, Mikhail B., Janion-Scheepers, Charlene, Deharveng, Louis (2020): Taxonomy of the Cryptopygus comp...

Published by Plazi.org taxonomic treatments database

... Taxonomy of the ***Cryptopygus*** complex. III. The ...

9 records

Taxonomy of the *Cryptopygus* complex. II. Affinity of austral *Cryptopygus* s. s. and Folsomia, with the description of two new Folsomia species (Collembola, Isotomidae)

Checklist dataset

Case study: Cryptopygus antarcticus

Cryptopygus antarcticus Willem, 1901	Antarctica	64.2S, 61.7W	2020 February	Human observation	iNat
Cryptopygus antarcticus Willem, 1901	Antarctica	62.8S, 61.4W	2020 January	Human observation	iNat
Cryptopygus antarcticus Willem, 1901	Norway		1978 February	Preserved specimen	Ento
Cryptopygus antarcticus Willem, 1901	Antarctica	64.5S, 62.3W		Preserved specimen	Type
Cryptopygus antarcticus Willem, 1901				Preserved specimen	NMN
Cryptopygus antarcticus Willem, 1901	Australia	41.9S, 146.1E	1992 April	Human observation	Tasr
Cryptopygus antarcticus Willem, 1901	Australia	41.9S, 146.1E	1992 April	Human observation	Tasr
Cryptopygus antarcticus Willem, 1901	Australia	42.2S, 146.1E	1989 January	Human observation	Tasr
Cryptopygus antarcticus Willem, 1901	Australia	42.8S, 146.6E	2001 February	Preserved specimen	Tasr

Misinterpretation of data standard

Missing information

Don’t forget down-stream cleaning of the data before use!!

Other biodiversity occurrence repositories

GBIF https://GBIF.org -> all biodiversity

OBIS OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM https://OBIS.org -> marine biodiversity (curated)

BIODIVERSITY.AQ www.biodiversity.aq -> Antarctic biodiversity (curated)

Nucleotide sequences

- DNA, RNA or proteins
- single genes or fragments
- Complete genomes
- Amplicon
  = PCR amplified fragments of a marker gene from environmental DNA
- Shotgun metagenomics
  = randomly fragmented environmental DNA
- Shotgun metatranscriptomics
  = randomly fragmented environmental RNA

Nucleotide sequences

- BLAST
  = query sequences based on small local alignments
- Standardized sequence data format: FASTA or FASTQ
- Standardized metadata format:

  Minimum Information on any (x) Sequence (MIxS)

  see webinar 2 (2020-11-03)

Nucleotide sequences

- Why look for nucleotide data?
  - Get reference genes or genomes
    - Comparative analyses
    - Gene prediction
    - Taxon identification
    - Phylogenetic analyses
  - Enrich data
    - More samples
    - cover more space
    - cover more time

= International Nucleotide Sequence Data Consortium

ENA

European Nucleotide Archive

NCBI

National Center for Biotechnology Information

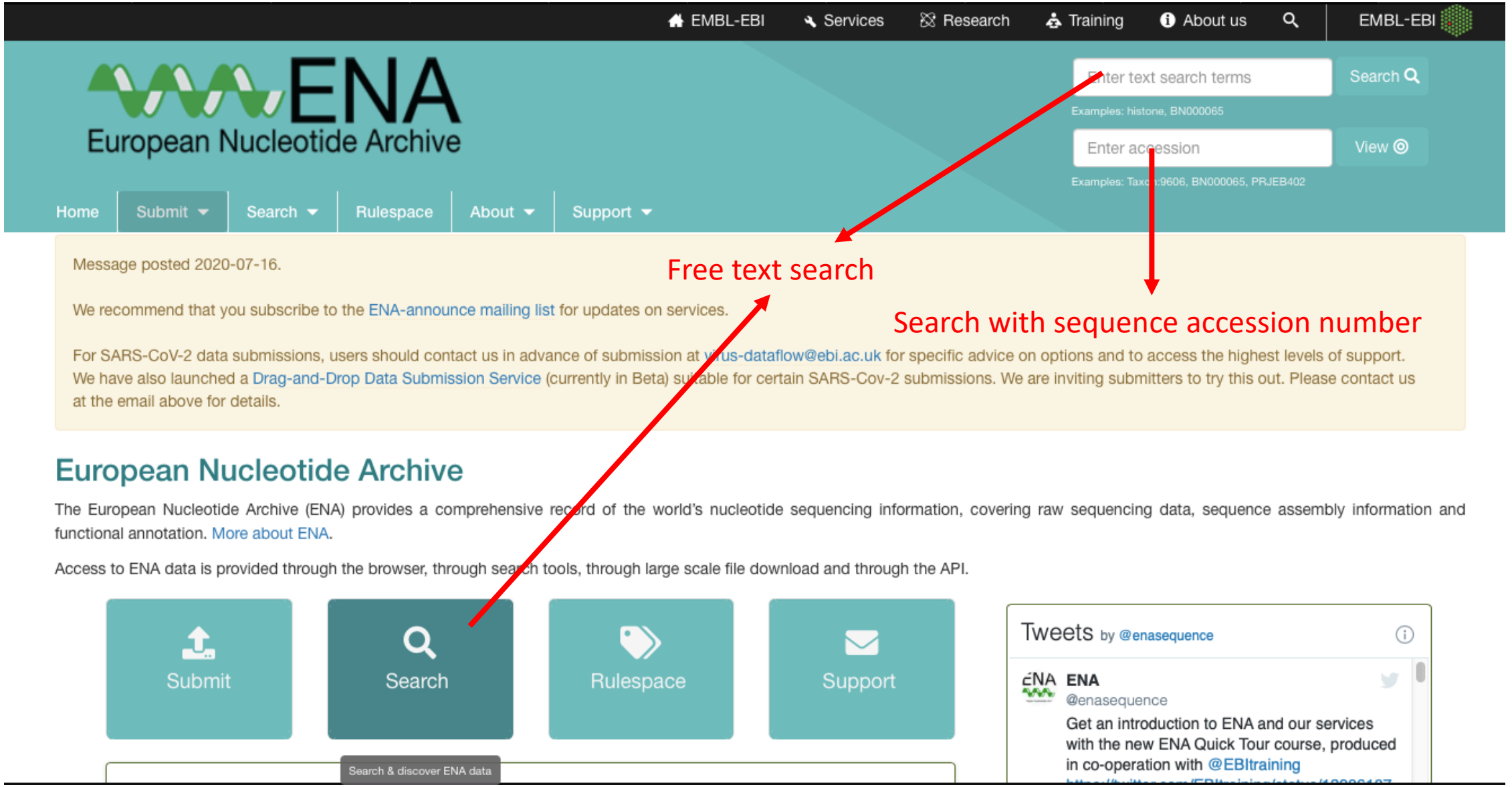
DDBJ

DNA Data Bank of Japan

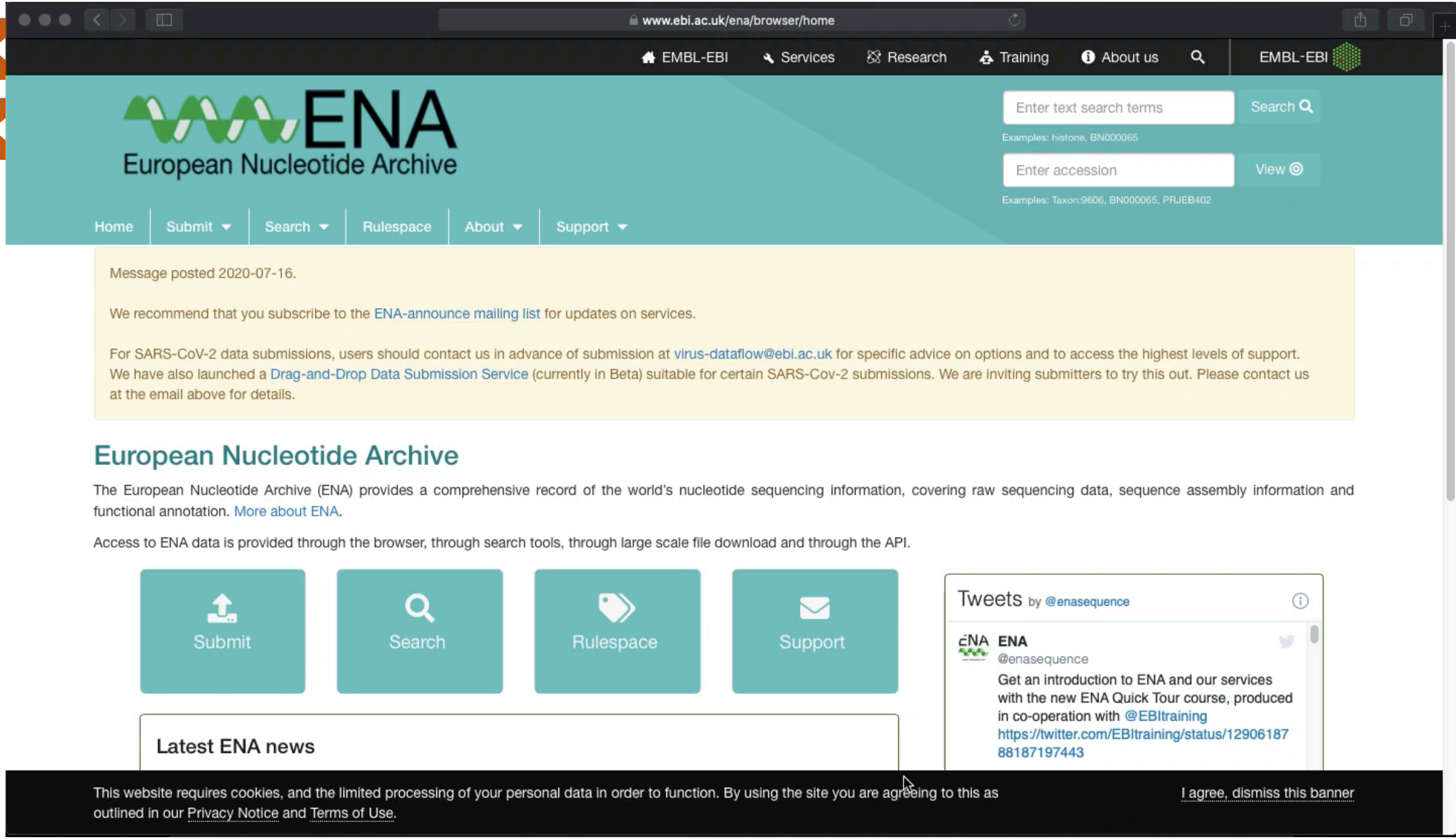
DEMO

example: Cryptopygus antarcticus

https://www.ebi.ac.uk/ena/browser/home



DEMO


www.ebi.ac.uk/ena/browser/home
EMBL-EBI
Services
Research
Training
About us
EMBL-EBI
ENA
European Nucleotide Archive
Enter text search terms
Search
Examples: histone, BN000065
Enter accession
View
Examples: Taxon:9606, BN000065, PRJEB402
Home
Submit
Search
Rulespace
About
Support
Message posted 2020-07-16.
We recommend that you subscribe to the ENA-announce mailing list for updates on services.
For SARS-CoV-2 data submissions, users should contact us in advance of submission at virus-dataflow@ebi.ac.uk for specific advice on options and to access the highest levels of support. We have also launched a Drag-and-Drop Data Submission Service (currently in Beta) suitable for certain SARS-Cov-2 submissions. We are inviting submitters to try this out. Please contact us at the email above for details.
European Nucleotide Archive
The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. More about ENA.
Access to ENA data is provided through the browser, through search tools, through large scale file download and through the API.
Submit
Search
Rulespace
Support
Tweets by @enasequence
ENA
@enasequence
Get an introduction to ENA and our services with the new ENA Quick Tour course, produced in co-operation with @EBItraining
https://twitter.com/EBItraining/status/1290618788187197443
Latest ENA news
This website requires cookies, and the limited processing of your personal data in order to function. By using the site you are agreeing to this as outlined in our Privacy Notice and Terms of Use.
I agree, dismiss this banner

DEMO

www.ebi.ac.uk/ena/browser/text-search?query=Cryptopygus%20antarcticus%2016S

ENA
European Nucleotide Archive

Cryptopygus antarcticus 16S | Search

Examples: histone, BN000065

Enter accession | View

Examples: Taxon:9606, BN000065, PRJEB402

Home | Submit | Search | Rulespace | About | Support

Text Search

Uses EBI Search to perform a free text search across ENA data. For more detailed usage please refer to the help & documentation section.

Search term: Cryptopygus antarcticus 16S | Search

Search results for Cryptopygus antarcticus 16S

- **Sequence**
  - Sequence (2)
  - Sequence (Standard) (2)
- **Non-coding**
  - Non-coding (2)
  - Non-coding (Standard) (2)

Sequence (Standard)

Download ENA records: FASTA TEXT

Accession	Description
MK433191	Cryptopygus antarcticus travei mitochondrion, complete genome.
EU016194	Cryptopygus antarcticus mitochondrion, complete genome.

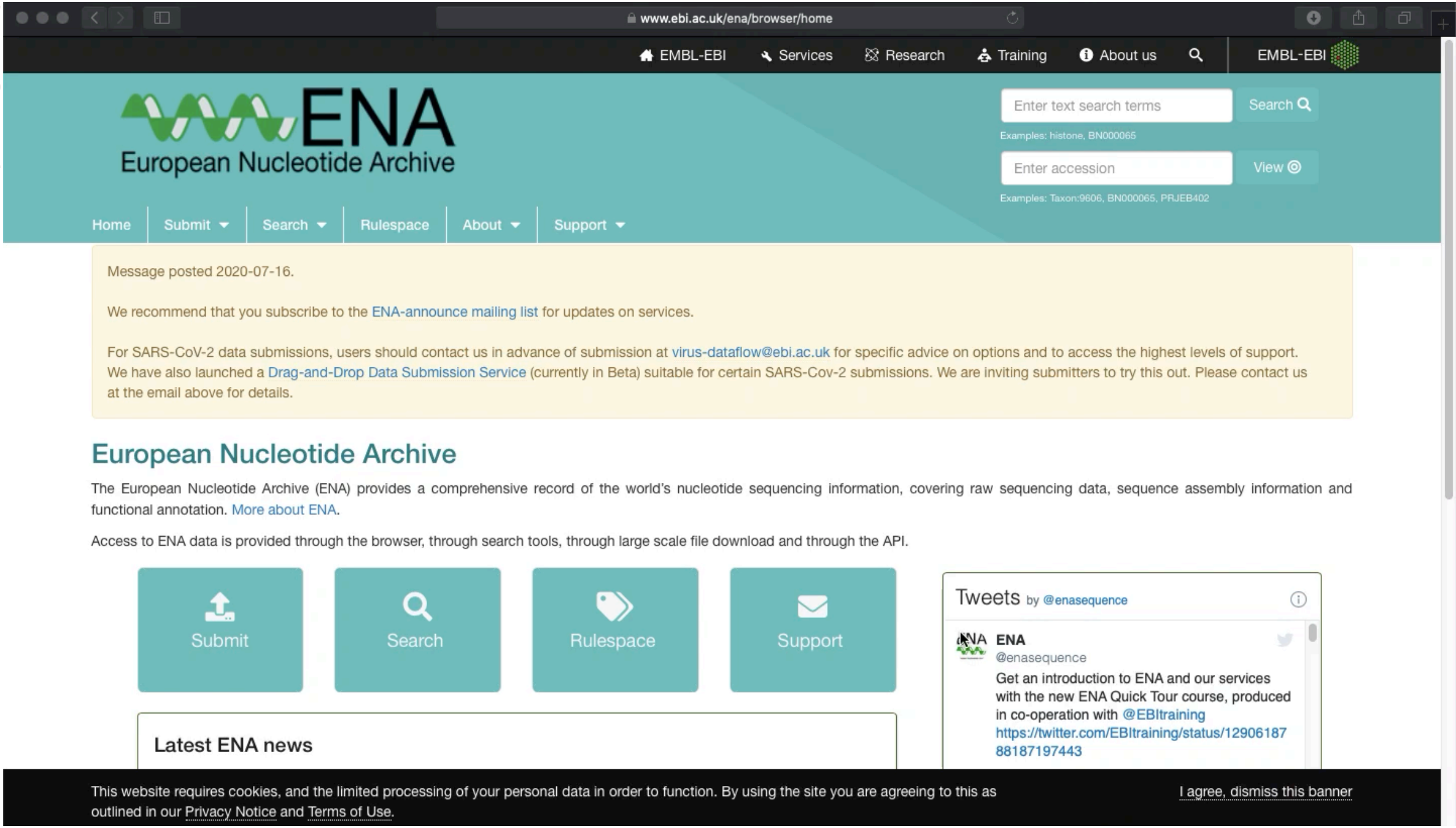
Items per page: 10 | 1 - 2 of 2

Powered by EBI Search

The European Nucleotide Archive (ENA) is part of the ELIXIR infrastructure

The ENA is an ELIXIR Core Data Resource. Learn more ›

DEMO


www.ebi.ac.uk/ena/browser/home
EMBL-EBI
Services
Research
Training
About us
EMBL-EBI
ENA
European Nucleotide Archive
Enter text search terms
Search
Examples: histone, BN000065
Enter accession
View
Examples: Taxon:9606, BN000065, PRJEB402
Home
Submit
Search
Rulespace
About
Support
Message posted 2020-07-16.
We recommend that you subscribe to the ENA-announce mailing list for updates on services.
For SARS-CoV-2 data submissions, users should contact us in advance of submission at virus-dataflow@ebi.ac.uk for specific advice on options and to access the highest levels of support. We have also launched a Drag-and-Drop Data Submission Service (currently in Beta) suitable for certain SARS-Cov-2 submissions. We are inviting submitters to try this out. Please contact us at the email above for details.
European Nucleotide Archive
The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. More about ENA.
Access to ENA data is provided through the browser, through search tools, through large scale file download and through the API.
Submit
Search
Rulespace
Support
Tweets by @enasequence
ENA
@enasequence
Get an introduction to ENA and our services with the new ENA Quick Tour course, produced in co-operation with @EBItraining
https://twitter.com/EBItraining/status/1290618788187197443
Latest ENA news
This website requires cookies, and the limited processing of your personal data in order to function. By using the site you are agreeing to this as outlined in our Privacy Notice and Terms of Use.
I agree, dismiss this banner

Where to find Nucleotide sequences

>90% of sequence data should be findable through INSDC databases:

- Genbank (NCBI)
- SRA (NCBI)
- ENA

- Other thematic databases for nucleotide data:

POLA3R https://www.biodiversity.aq/pola3r/ -> polar regions

BOLD SYSTEMS https://boldsystems.org/ -> barcode of life

unite community https://unite.ut.ee/ -> barcodes microbial eukaryotes

Search POLA$_3$R by:

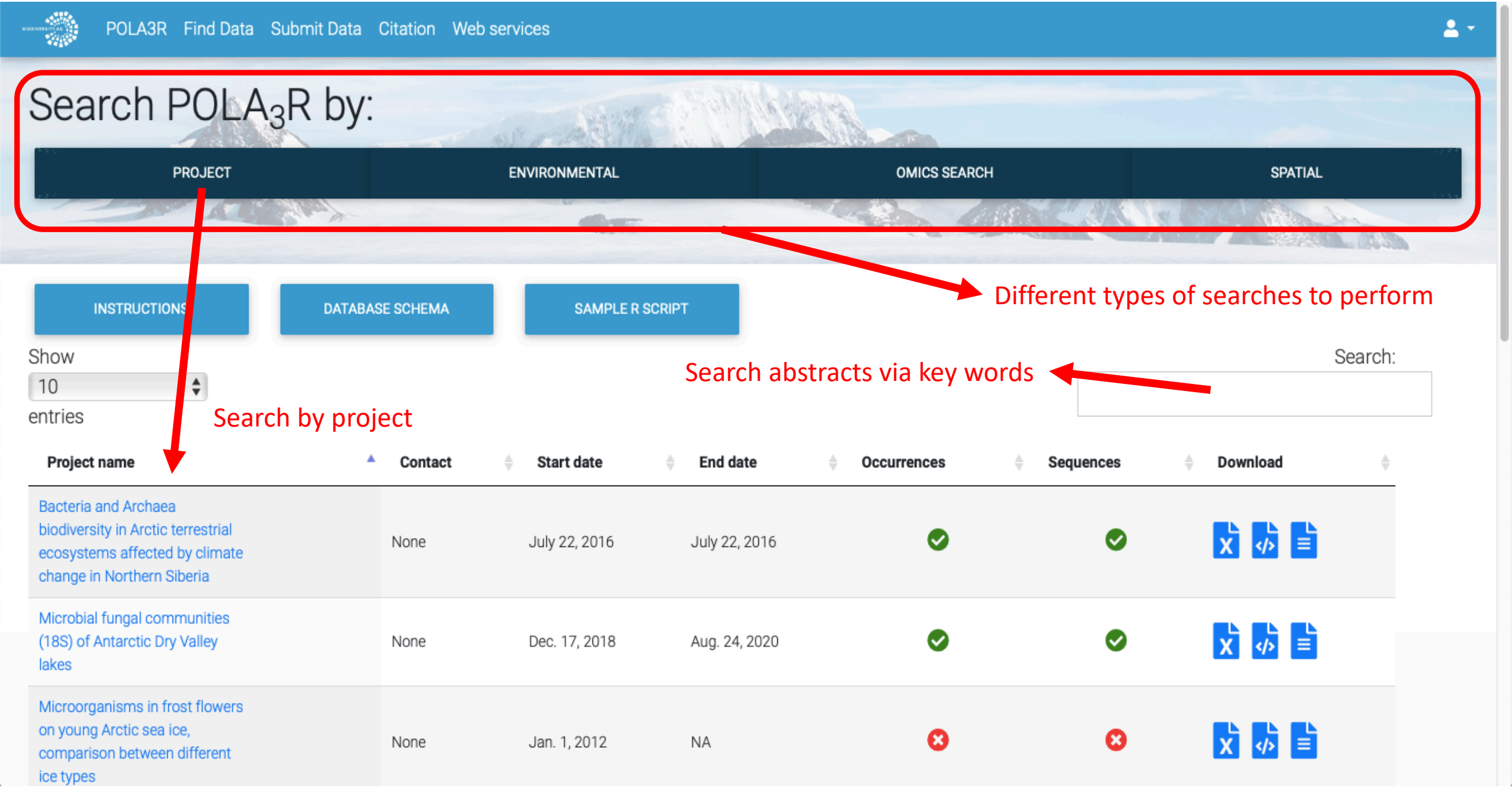
PROJECT | ENVIRONMENTAL | OMICS SEARCH | SPATIAL

INSTRUCTIONS | DATABASE SCHEMA | SAMPLE R SCRIPT

Show 10 entries

Search:

Project name	Contact	Start date	End date	Occurrences	Sequences	Download
Bacteria and Archaea biodiversity in Arctic terrestrial ecosystems affected by climate change in Northern Siberia	None	July 22, 2016	July 22, 2016	✓	✓	
Microbial fungal communities (18S) of Antarctic Dry Valley lakes	None	Dec. 17, 2018	Aug. 24, 2020	✓	✓	
Microorganisms in frost flowers on young Arctic sea ice, comparison between different ice types	None	Jan. 1, 2012	NA	✗	✗	


POLA3R Find Data Submit Data Citation Web services
Search POLA3R by:
PROJECT
ENVIRONMENTAL
OMICS SEARCH
SPATIAL
INSTRUCTIONS
DATABASE SCHEMA
SAMPLE R SCRIPT
Different types of searches to perform
Show
10
entries
Search by project
Search abstracts via key words
Search:
Project name
Contact
Start date
End date
Occurrences
Sequences
Download
Bacteria and Archaea biodiversity in Arctic terrestrial ecosystems affected by climate change in Northern Siberia
None
July 22, 2016
July 22, 2016
Microbial fungal communities (18S) of Antarctic Dry Valley lakes
None
Dec. 17, 2018
Aug. 24, 2020
Microorganisms in frost flowers on young Arctic sea ice, comparison between different ice types
None
Jan. 1, 2012
NA

Environmental data

- In ecology
  - Correlations of biodiversity or community composition
- Many (inter-) national monitoring scheme’s available
- Climate and geophysical
  - Ground measurements (e.g. weather stations)
  - Remote sensing, satellite (e.g. NASA)
- Co-recorded with biodiversity/nucleotide data
  - Environmental measurements (e.g. pH, conductivity,...)
  - Usually associated with biodiversity data
  - Often difficult to find...

Ocean Data

NOAA

World Ocean Atlas

https://www.nodc.noaa.gov/OC5/woa18/ -> ocean data (USA)

https://www.bio-oracle.org -> layers for ecological modeling

EMODnet

https://emodnet.eu/en -> ocean data (EU)

Environmental data

- Other examples:

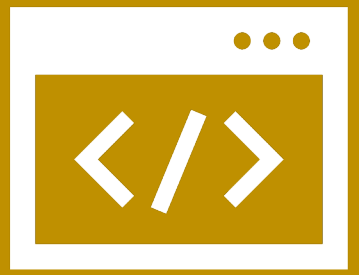
https://www.nccs.nasa.gov/ -> NASA remote sensing and climate

https://copernicus.eu -> ocean remote sensing and climate (EU)

https://www.worldclim.org/data/bioclim.html
-> BioClim variable (precipitation and temperature)

https://developers.google.com/earth-engine/datasets
-> datalayers for Google Earth

https://www.polarview.aq/antarctic -> Antarctic data

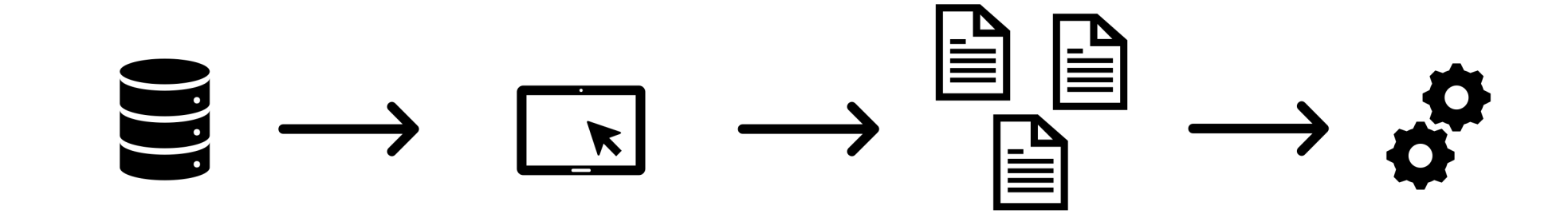


Advanced ways to get data

- Big data = automate

- Tools of the trade:
  - Functions
    = define an action to a computer
  - For-loops
    = let a computer repeat an action
    (while you do something else)
  - Web services
    = let a computer interact directly with a website

Web services



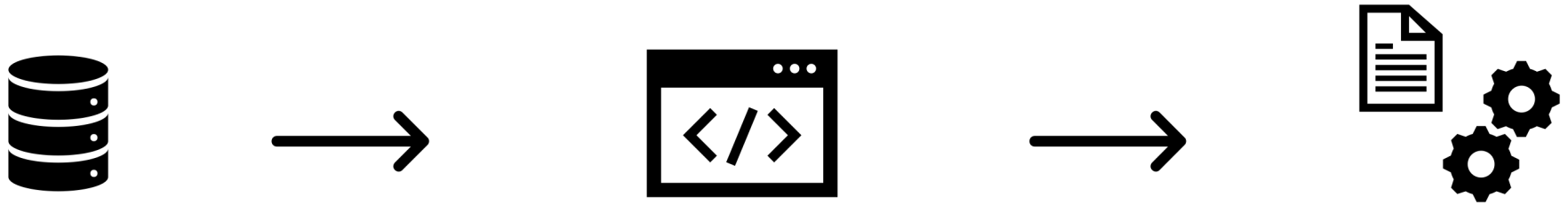
Back-end database

Front-end interface
(= click and point web pages)

Individual data downloads

Analysis pipeline

Web services



Back-end database

Direct queries using an
Application Programming Interface (API)
(=web services)
(= send requests through URLs)

Download all data directly into
analysis pipeline

- Example of our case study:

  https://api.gbif.org/v1/occurrence/search?scientificName=Cryptopygus antarcticus

- Results can be downloaded and processed directly into a pipeline

DEVELOPER | API DOCS

API Summary

https://api.gbif.org/v1/

The GBIF API is a RESTful JSON based API. The base URL for v1 you should use is: **https://api.gbif.org/v1/**

The API should be considered stable, as should this accompanying documentation. It is also available with HTTPS. Please report any issues you find with either the API itself or the documentation using the "feedback" button on the top right.

Content

API Sections

The API is split into logical sections to ease understanding:

Registry: Provides means to create, edit, update and search for information about the datasets, organizations (e.g. data publishers), networks and the means to access them (technical endpoints). The registered content controls what is crawled and indexed in the GBIF data portal, but as a shared API may also be used for other initiatives
Species: Provides services to discover and access information about species and higher taxa, and utility services for interpreting names and looking up the identifiers and complete scientific names used for species in the GBIF portal.
Occurrence: Provides access to occurrence information crawled and indexed by GBIF and search services to do real time paged search and asynchronous download services to do large batch downloads.
Maps: Provides simple services to show the maps of GBIF mobilized content on other sites.
News: Provides services to stream useful information such as papers published using GBIF mobilized content for various themes.

Computer languages as tools

+ Common for statistics and sciences
+ big community base (packages)
+ Good with data
- Slow for really big data

+ all sorts of applications
+ Easy syntax
- less possibilities for data analysis

+ data manipulation
+ interact with the web
- niche user base
- not suited for data analysis

Case study: occurrence data in R

Freely available at https://rstudio.com

rgbif package: communicate with GBIF from R over the WWW

Code repo: https://github.com/ropensci/rgbif

example: occurrence data in R

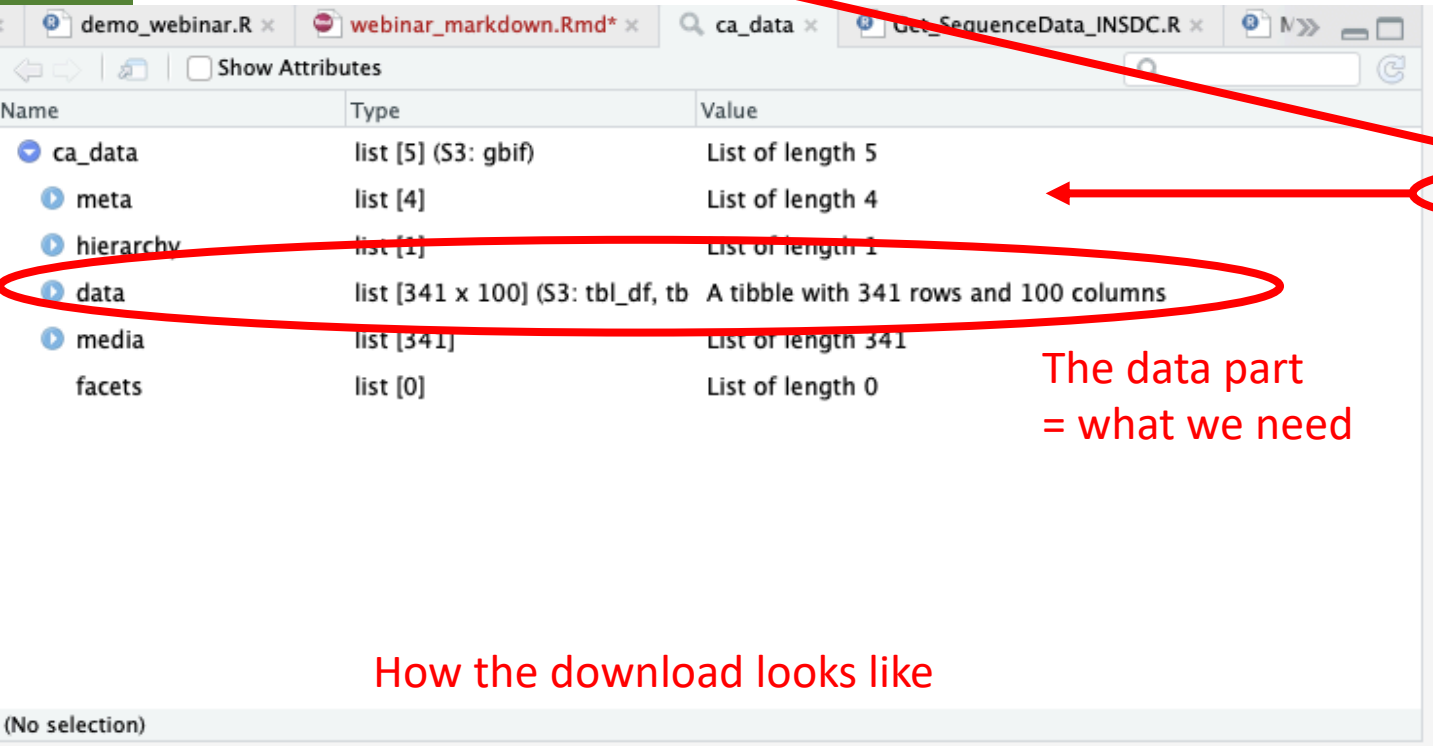
```
# install the rgbif package
# a package = code that enables to do certain things (functions)
install.packages("rgbif")
```

```
# load the rgbif library (=make the code available to R)
library(rgbif)

# set the working directory
# this is where R will look for data or send data when you save
setwd("/User/file/file_where_my_scripts_and_data_goes")
```

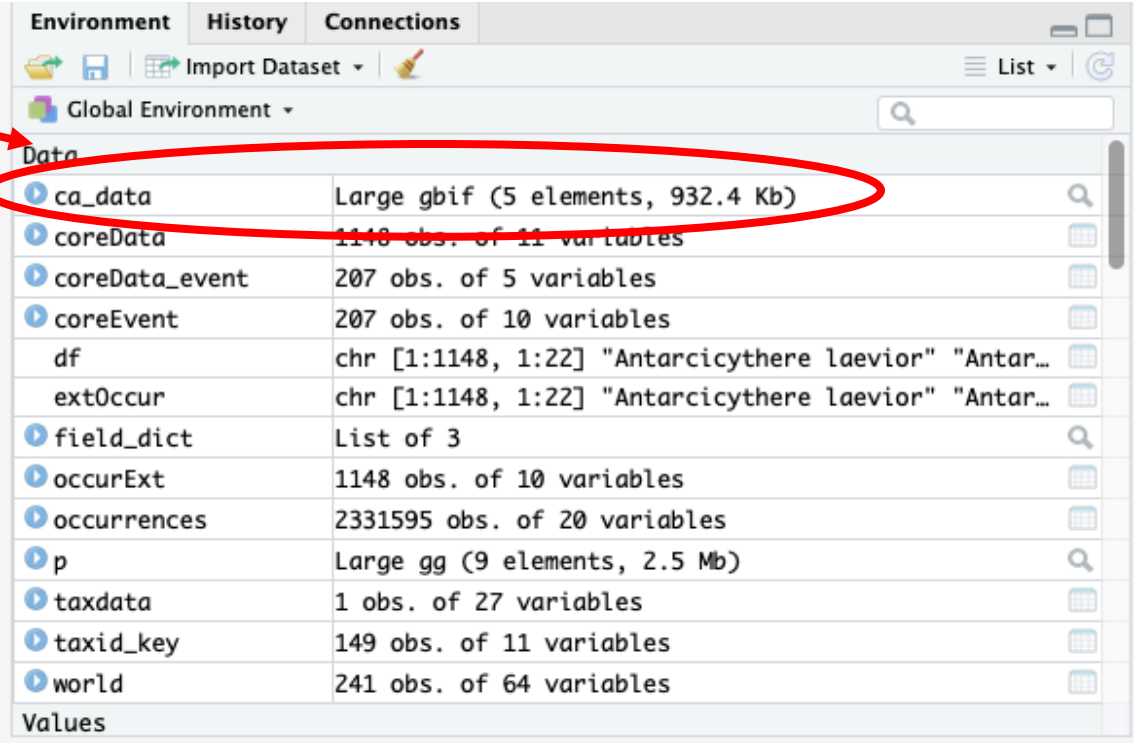
example: occurrence data in R

```
# query GBIF and download occurrences
# look for occurrences of our species
ca_data <- rgbif::occ_search(scientificName= "Cryptopygus antarcticus")
```



The data part
= what we need

How the download looks like



example: occurrence data in R

```
# query GBIF and download occurrences
# look for occurrences of our species
ca_data <- rgbif::occ_search(scientificName= "Cryptopygus antarcticus")
```

```
ca_data <- data.frame(ca_data$data)
```

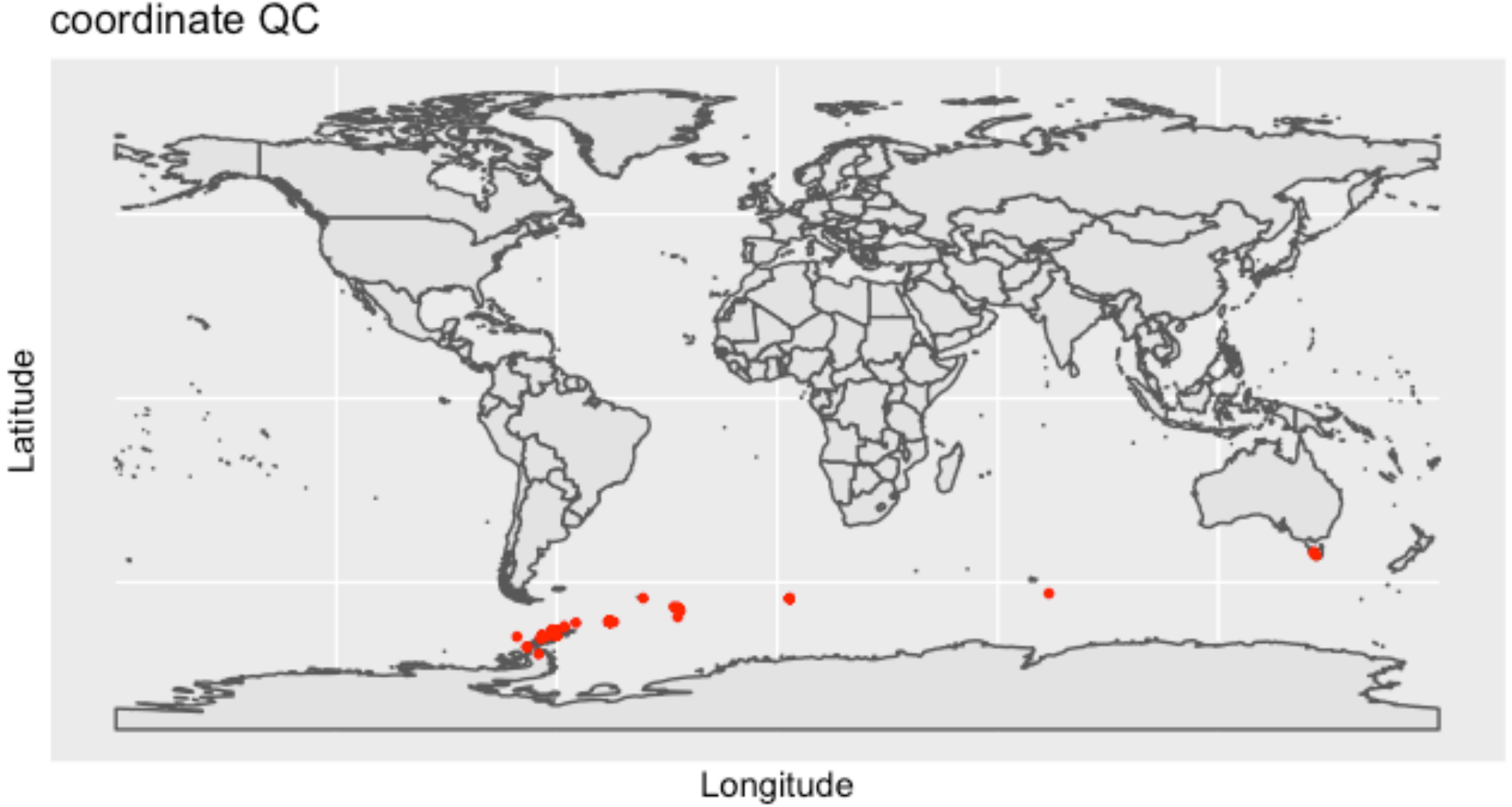
example: occurrence data in R

```
# plot the occurrences on a workd map
library(ggplot2) # a popular library to make fancy graphs and images
library(rnaturalearth) # the data of the world map (country outline polygons)

# collect the data for the map plot at the prefered resolution
world <- ne_countries(scale = "medium", returnclass = "sf")

# plot the occurrences based on their decimal coordinates
p<- ggplot(data = world) +
      geom_sf() +
      xlab("Longitude") + ylab("Latitude") +
      ggtitle("coordinate QC") +
      geom_point(data = ca_data, aes(y=decimalLatitude, x=decimalLongitude),
                 colour="red",size=0.9)
print(p)
```

example: occurrence data in R



example: occurrence data in R

- Further data cleaning before analysis is a must!
  - Check the geographic coordinates
  - Exclude data based on completeness (date, place,...)
  - Sort out wrongly formatted data (e.g. non ISO date-time)
  - ….
- Decide exclusion criteria before hand

example: occurrence data in R

- Other ways of searching:

```
?occ_search
```

example: occurrence data in R

- Othe

```
?occ_se
```

Search for GBIF occurrences

Description

Search for GBIF occurrences

Usage

```
occ_search(
  taxonKey = NULL,
  scientificName = NULL,
  country = NULL,
  publishingCountry = NULL,
  hasCoordinate = NULL,
  typeStatus = NULL,
  recordNumber = NULL,
  lastInterpreted = NULL,
  continent = NULL,
  geometry = NULL,
  geom_big = "asis",
  geom_size = 40,
  geom_n = 10,
  recordedBy = NULL,
  recordedByID = NULL,
  identifiedByID = NULL,
  basisOfRecord = NULL,
  datasetKey = NULL,
  eventDate = NULL,
  catalogNumber = NULL,
```

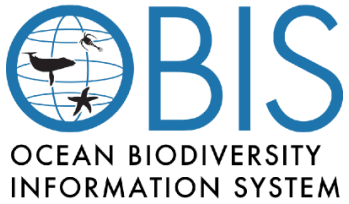
example: occurrence data in R

- Other R tools for occurrence data:

rgbif package: communicate with GBIF from R over the WWW

Code repo: https://github.com/ropensci/rgbif

robis and spocc



robis package: communicate with OBIS from R

Code repo: https://github.com/iobis/robis

spocc package: An interface to many species occurrence data sources (GBIF, OBIS, BISON, iNaturalist, eBird,...

Code repo: https://github.com/ropensci/spocc

Spocc images: phylopic.org

For-loop

- Syntax

for(number of iterations){ action }

```
for(i in 1:10){
        print(i)
}
```

1:10 = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

action = in each iteration, print the value of “i”

For-loop

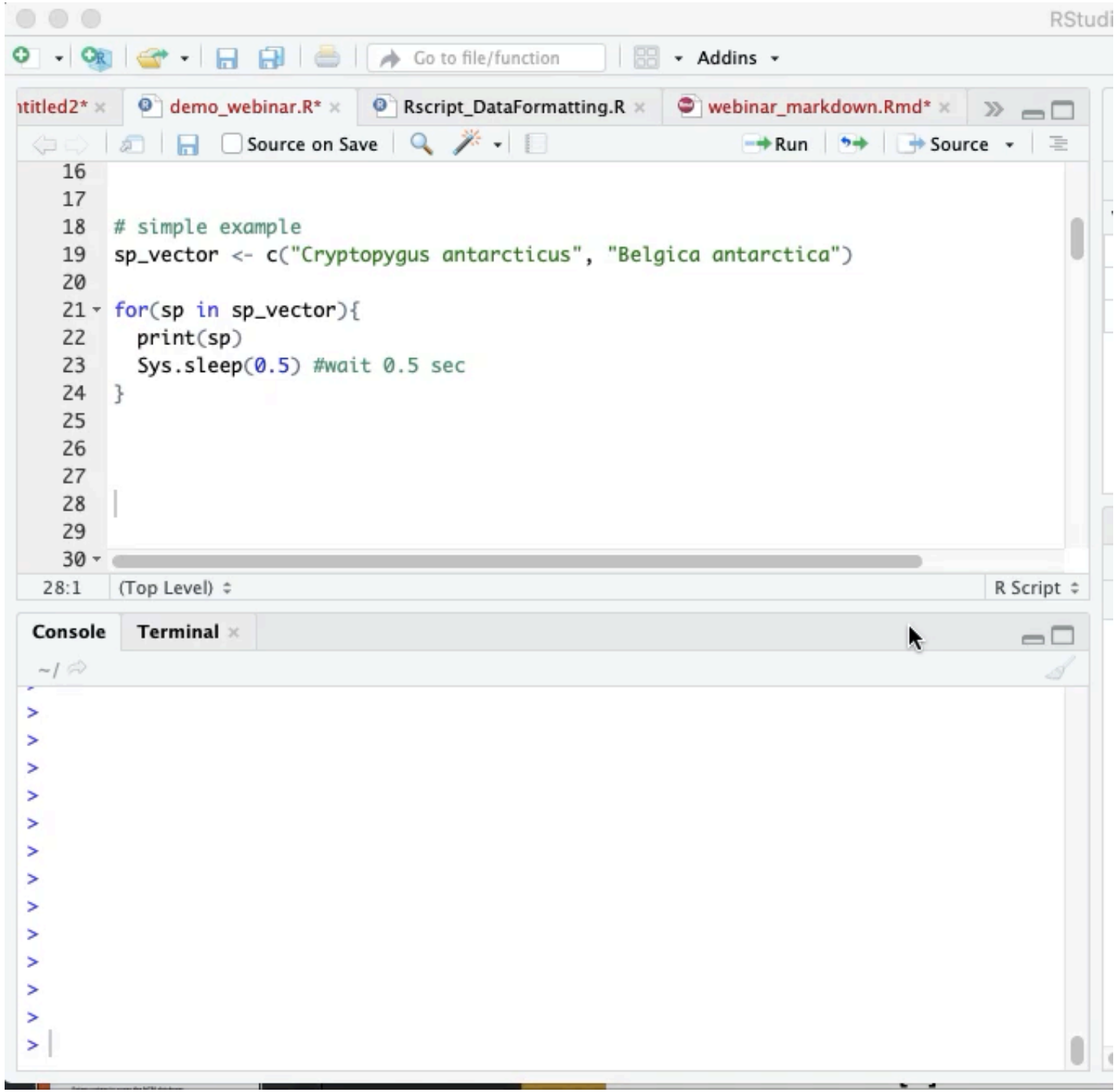
```
for(i in 1:10){
        print(i)
}
```

```
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
[1] 6
[1] 7
[1] 8
[1] 9
[1] 10
```

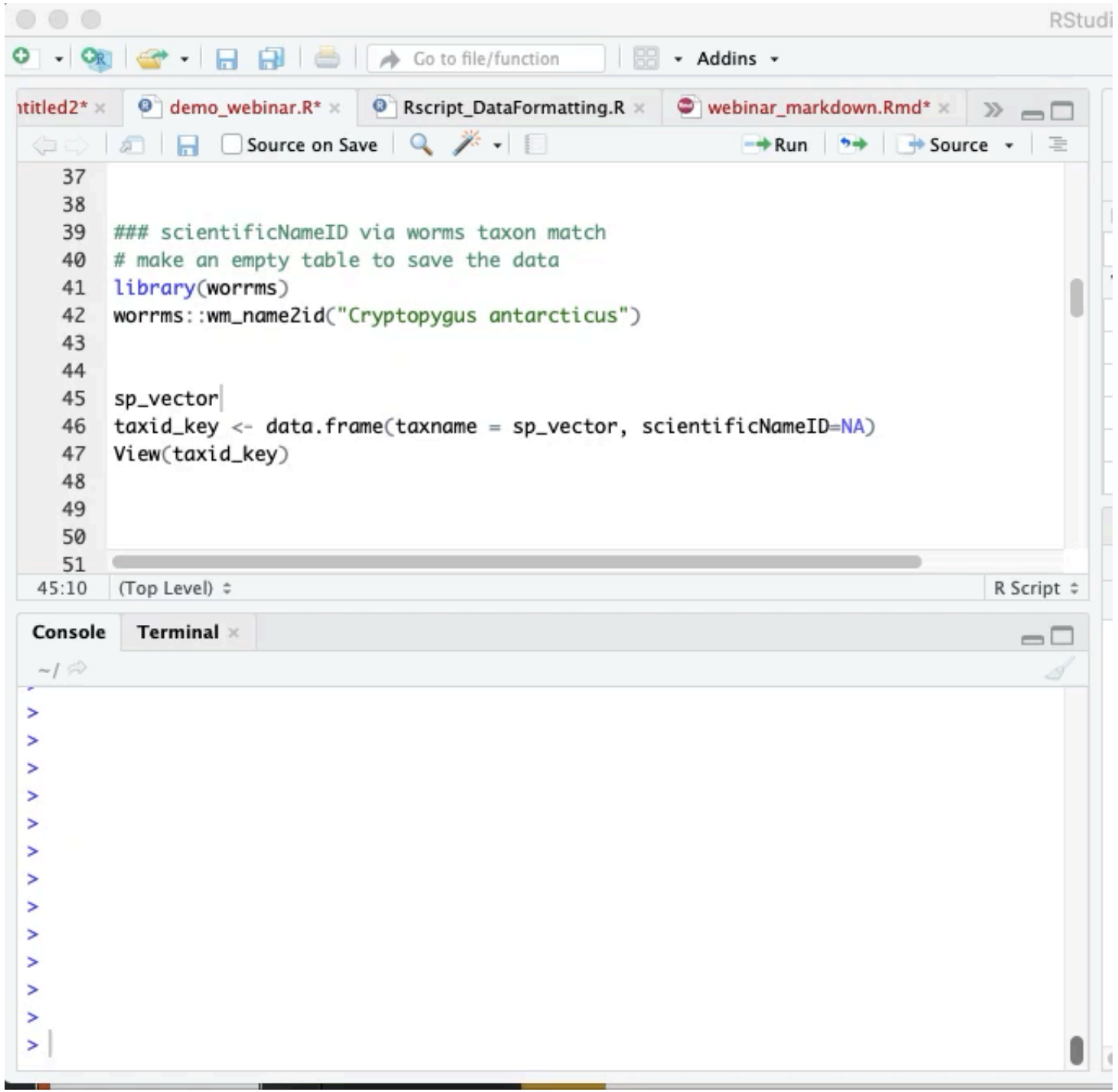
Result of the print function during iteration 1

iteration 2

For-loop


RStudi
Go to file/function
Addins
untitled2*
demo_webinar.R*
Rscript_DataFormatting.R
webinar_markdown.Rmd*
Source on Save
Run
Source
16
17
18 # simple example
19 sp_vector <- c("Cryptopygus antarcticus", "Belgica antarctica")
20
21 for(sp in sp_vector){
22 print(sp)
23 Sys.sleep(0.5) #wait 0.5 sec
24 }
25
26
27
28
29
30
28:1
(Top Level)
R Script
Console
Terminal
~/

For-loop


RStudi
Go to file/function
Addins
titled2*
demo_webinar.R*
Rscript_DataFormatting.R
webinar_markdown.Rmd*
Source on Save
Run
Source
37
38
39 ### scientificNameID via worms taxon match
40 # make an empty table to save the data
41 library(worrms)
42 worrms::wm_name2id("Cryptopygus antarcticus")
43
44
45 sp_vector
46 taxid_key <- data.frame(taxname = sp_vector, scientificNameID=NA)
47 View(taxid_key)
48
49
50
51
45:10
(Top Level)
R Script
Console
Terminal
~/

example: sequence data in R

MicrobeDataTools package: tools to download, process and archive environmental sequence data

Code repo: https://github.com/biodiversity-aq/MicrobeDataTools

- INSDC makes data available through a REST API called the ”Entrez Utilities” (Eutils)

  =a set of server-side programs for stable interface to the

  Entrez system to query the NCBI databases

- Data can be downloaded from an FTP server

See also: the “rentrez” package

Case study: sequence data through R

```
library(devtools) #a package specialiced in development tools, allows to install
from GitHub
Sys.setenv(R_REMOTES_NO_ERRORS_FROM_WARNINGS="true") #default false gives error
remotes::install_github("biodiversity-aq/MicrobeDataTools", dependencies=FALSE)
```

- MicrobeDataTools is still in development, if it doesn't install properly you can find all the code in the GitHub repo.

Case study: sequence data through R

```
# load the package
library(MicrobeDataTools)
# set the working directory
# this is where R will look for data or send data when you save
setwd("/User/file/file_where_my_scripts_and_data_goes")
```

Case study: sequence data through R

```
api_personalKeyMaxime <- "************************************"

# download sequences from INSDC to R
# this requires registration of a personal API key to INSDC
# same as downloading from the site, but directly to a file
run_meta <- MicrobeDataTools::download.sequences.INSDC(BioPrj =
c("PRJNA397058"), apiKey=api_personalKeyMaxime)
```

- This requires an API key
- = unique string that you include in your HTTP requests that identifies you to INSDC servers
- Can be requested at NCBI: https://www.ncbi.nlm.nih.gov/account/

```
> run_meta <- MicrobeDataTools::download.sequences.INSDC(BioPrj = c("PRJNA397058"), apiKey=api_personalKeyMaxime)
Notice!
the metadata will be retruned to the Console
If you did assign the output of this function to an R-object (using "<-"): better abort and restart now


Getting the metadata ...

Processing BioProject PRJNA397058...

        43 samples (Runs)...

        processing the metadata ...

        Downloading the sequence data ...

trying URL 'ftp.sra.ebi.ac.uk/vol1/fastq/SRR589/000/SRR5894260/SRR5894260.fastq.gz'
Content type 'unknown' length 53465 bytes (52 KB)
==================================================
trying URL 'ftp.sra.ebi.ac.uk/vol1/fastq/SRR589/003/SRR5894243/SRR5894243.fastq.gz'
Content type 'unknown' length 31550 bytes (30 KB)
==================================================
trying URL 'ftp.sra.ebi.ac.uk/vol1/fastq/SRR589/008/SRR5894248/SRR5894248.fastq.gz'
```

Case study: sequence data through R

```
# the data in "run_meta" is the metadata of each sample
# this is is not present when downloading the sequences from INSDC
# the metadata can also be downloaded separately:
# some basic metadata can be downloaded without API authentification
basic_meta <- MicrobeDataTools::get.BioProject.metadata.INSDC("PRJNA397058")

# this does not include environmental measurements or coordinates
#these can be downloaded with the following function
full_meta <- MicrobeDataTools::get.sample.attributes.INSDC(BioPrjct="PRJNA397058
", apiKey= api_personalKeyMaxime)
# have a look at the data
View(full_meta)
```

Case study: sequence data through R

```
# the data in "run_meta" is the metadata of each sample
# this is is not present when downloading the sequences from INSDC
```

demo_webinar.R | webinar_markdown.Rmd* | full_meta | Get_SequenceData_INSDC.R | MAKE_MicrobeDataTools.R | POLAAAR_polulateDB_12oct2020.R

Filter

	attr_name	BioSampleModel	collection_date	geo_loc_name	isolation_source	lat_lon	BioProject	SRA_sample	LibraryName
SRR5894242	SRR5894242	Metagenome or environmental	15-Nov-2014	Antarctica	Sedimentary rock Union Glacier T13	79.74 S 83.21 W	PRJNA397058	SRS2405776	DecepM1
SRR5894243	SRR5894243	Metagenome or environmental	15-Jan-2014	Antarctica	Soil Deception M6	62.98 S 60.54 W	PRJNA397058	SRS2405775	DecepM2
SRR5894244	SRR5894244	Metagenome or environmental	15-Jan-2014	Antarctica	Soil Deception M5	62.99 S 60.54 W	PRJNA397058	SRS2405778	DecepM3
SRR5894245	SRR5894245	Metagenome or environmental	15-Jan-2014	Antarctica	Soil Deception M8	62.98 S 60.54 W	PRJNA397058	SRS2405777	DecepM4
SRR5894246	SRR5894246	Metagenome or environmental	15-Jan-2014	Antarctica	Soil Deception M7	62.98 S 60.53 W	PRJNA397058	SRS2405772	DecepM5
SRR5894247	SRR5894247	Metagenome or environmental	15-Jan-2014	Antarctica	Soil Deception M2	62.98 S 60.66 W	PRJNA397058	SRS2405771	DecepM6
SRR5894248	SRR5894248	Metagenome or environmental	15-Jan-2014	Antarctica	Soil Deception M1	62.98 S 60.70 W	PRJNA397058	SRS2405774	DecepM7
SRR5894249	SRR5894249	Metagenome or environmental	15-Jan-2014	Antarctica	Soil Deception M4	62.99 S 60.68 W	PRJNA397058	SRS2405773	DecepM8
SRR5894250	SRR5894250	Metagenome or environmental	15-Jan-2014	Antarctica	Soil Deception M3	62.99 S 60.67 W	PRJNA397058	SRS2405795	DeeM14
SRR5894251	SRR5894251	Metagenome or environmental	15-Jan-2014	Antarctica	Soil Snow M10	62.73 S 61.23 W	PRJNA397058	SRS2405794	DeeM15
SRR5894252	SRR5894252	Metagenome or environmental	15-Jan-2014	Antarctica	Soil Snow M9	62.73 S 61.22 W	PRJNA397058	SRS2405793	DeeM16
SRR5894253	SRR5894253	Metagenome or environmental	15-Nov-2014	Antarctica	Sedimentary rock Union Glacier T10	79.73 S 83.17 W	PRJNA397058	SRS2405792	DeeM17
SRR5894254	SRR5894254	Metagenome or environmental	15-Nov-2014	Antarctica	Sedimentary rock Union Glacier T9	79.73 S 83.17 W	PRJNA397058	SRS2405791	GreenM18
SRR5894255	SRR5894255	Metagenome or environmental	15-Nov-2014	Antarctica	Sedimentary rock Union Glacier T5	79.81 S 83.27 W	PRJNA397058	SRS2405808	LagI1
SRR5894256	SRR5894256	Metagenome or environmental	15-Nov-2014	Antarctica	Sedimentary rock Union Glacier T3	79.79 S 82.94 W	PRJNA397058	SRS2405810	LagI2

Showing 1 to 16 of 43 entries

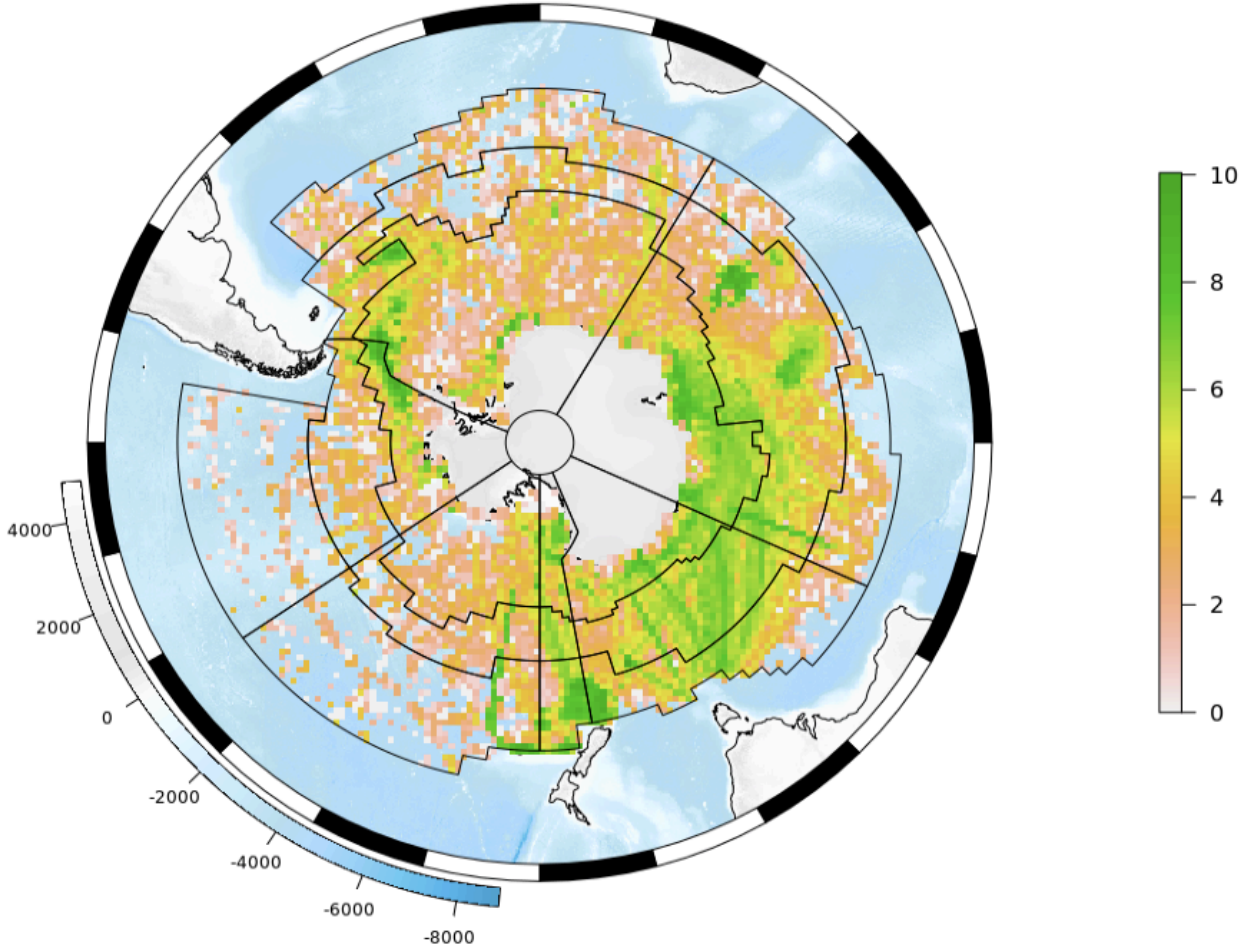
Case study: sequence data through R

- downstream processing of the sequences
  - e.g. the DADA2 pipeline (Callahan et al. 2016; doi: 10.1038/nmeth.3869) https://benjjneb.github.io/dada2/tutorial.html
  - Cluster in Alternative Sequence Variants (ASVs)
  - Compare to a curated database (e.g. PR2 for 18S rRNA)

Environmental data

SOmap package: plotting maps of Antarctica

Code repo: https://github.com/AustralianAntarcticDivision/SOmap



Environmental data

blueant package: retrieving environmental information

Code repo: https://github.com/AustralianAntarcticDivision/blueant

SOhungry package: Southern Ocean diet and energetics data

Code repo: https://github.com/SCAR/sohungry

Final remarks

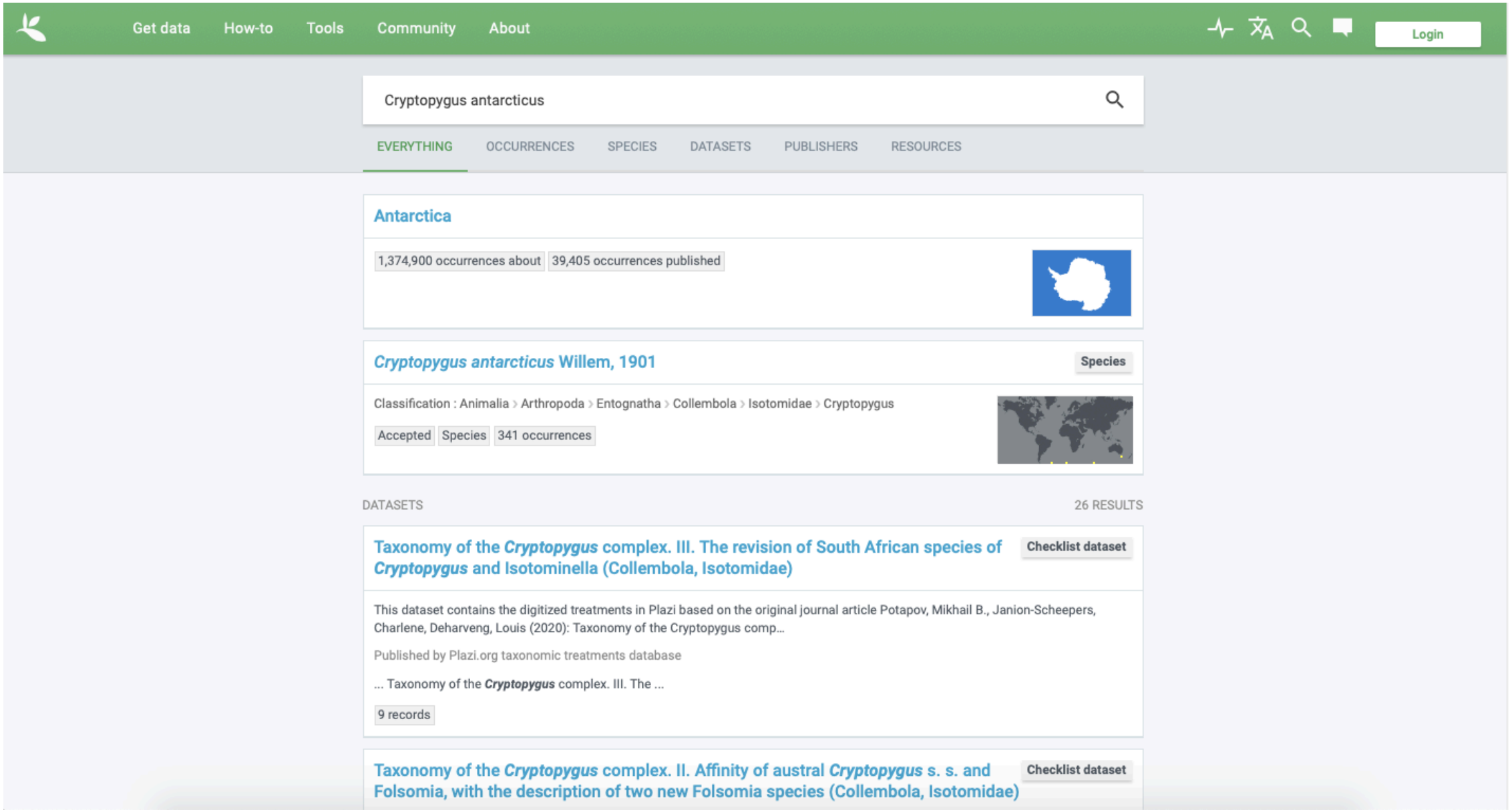
- Before you start:
  - set appropriate exclusion/inclusion criteria
  - choose the search terms
- Sharing is caring
  - Make your data public as well
  - Make your code available
- Legal
  - Find the appropriate citation
  - Check the license of use
  - Be transparent on how data was obtained and processed

Thanks for your attention

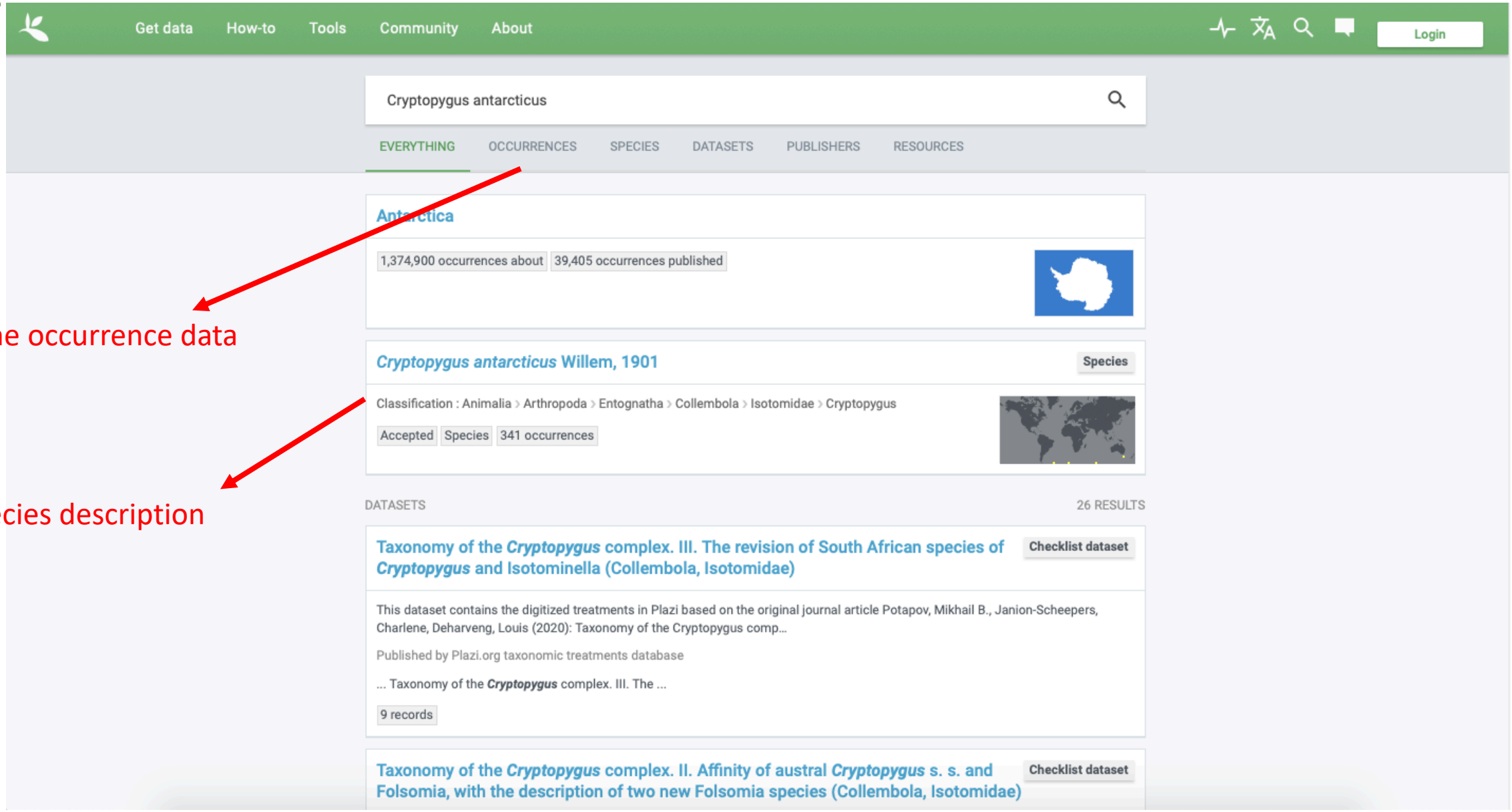
Thanks for your attention

Backup if movies fail:

Case study: Cryptopygus antarcticus



Case study: Cryptopygus antarcticus



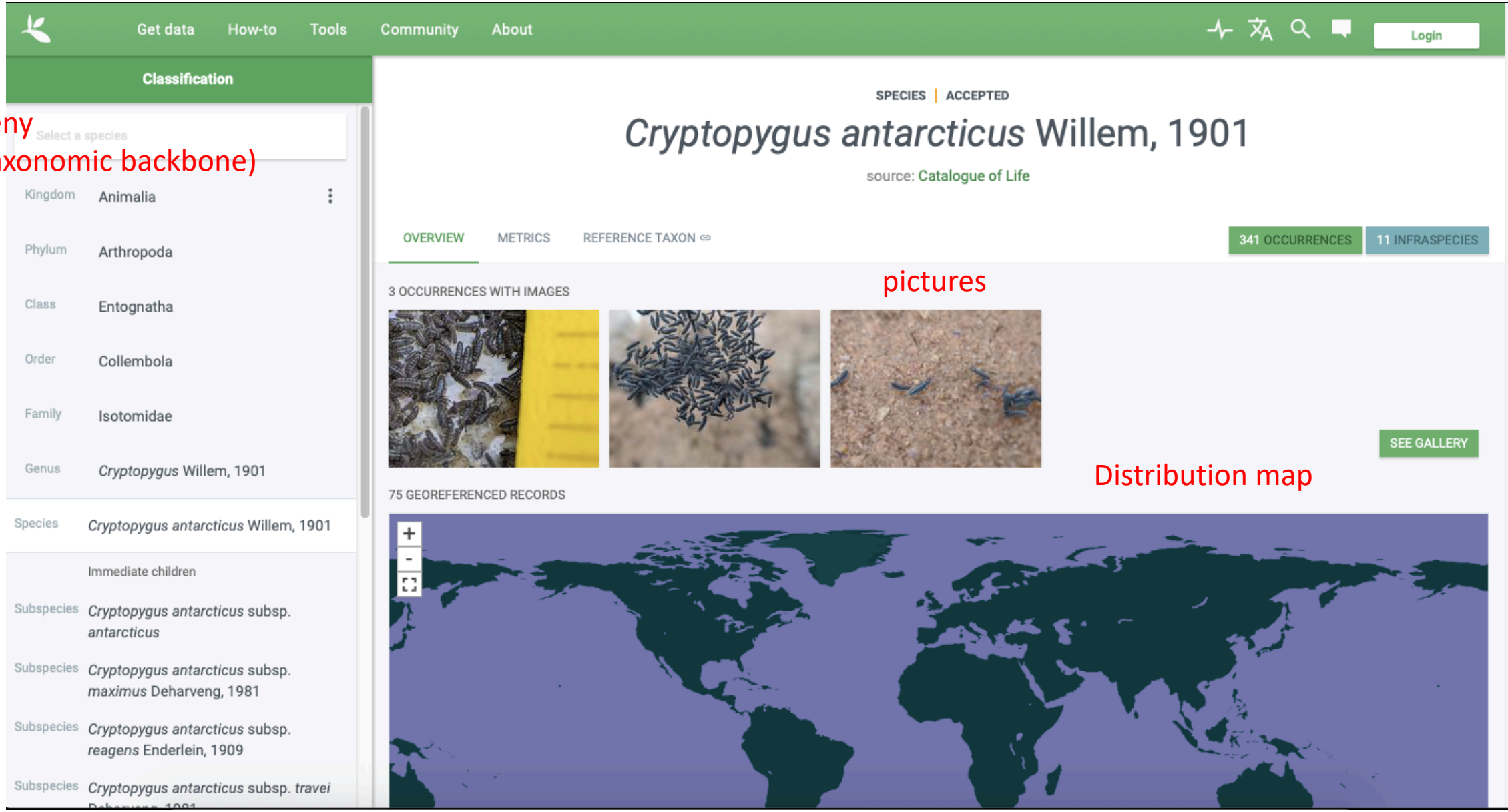
To the occurrence data

Species description

DEMO

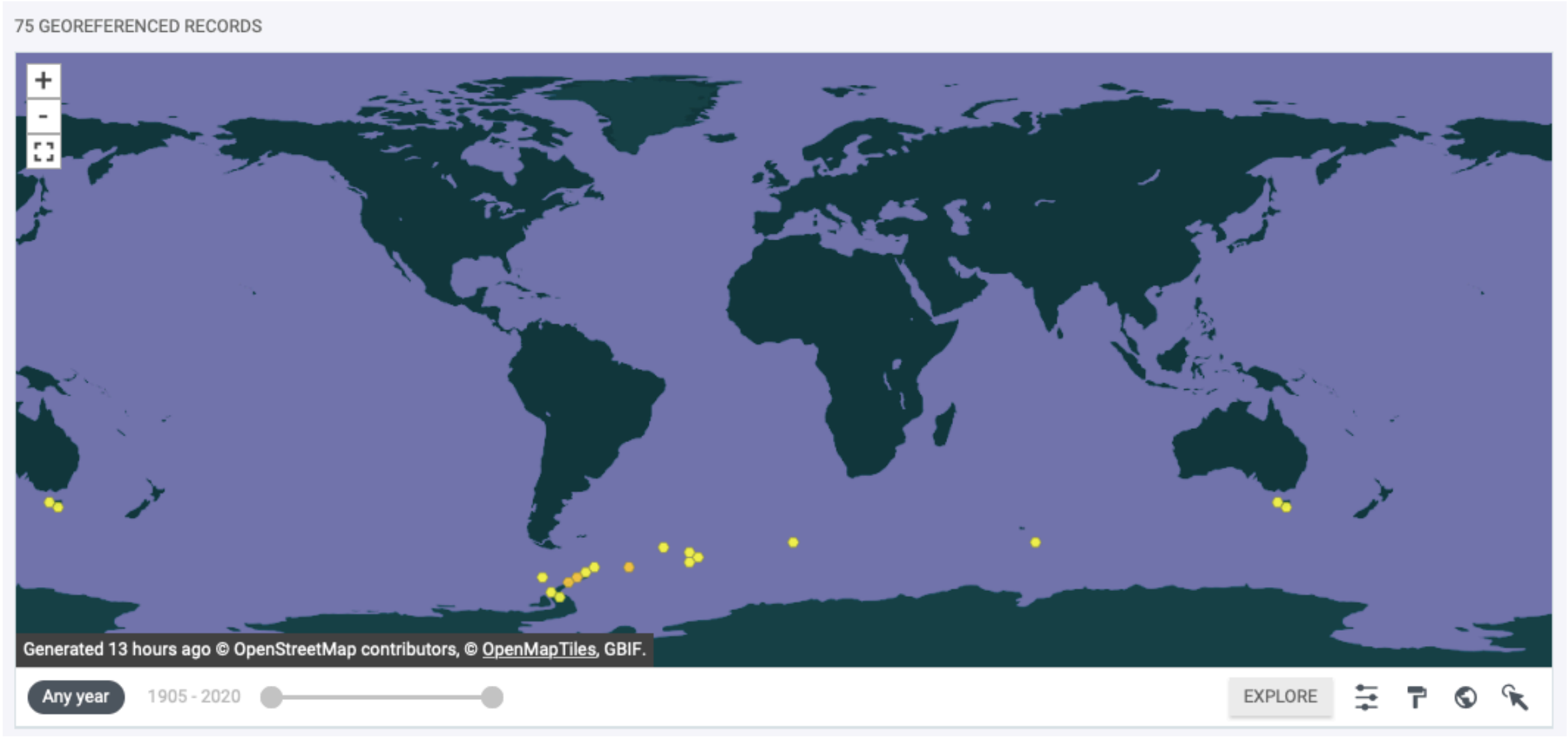
Case study: Cryptopygus antarcticus

Phylogeny
(GBIF taxonomic backbone)



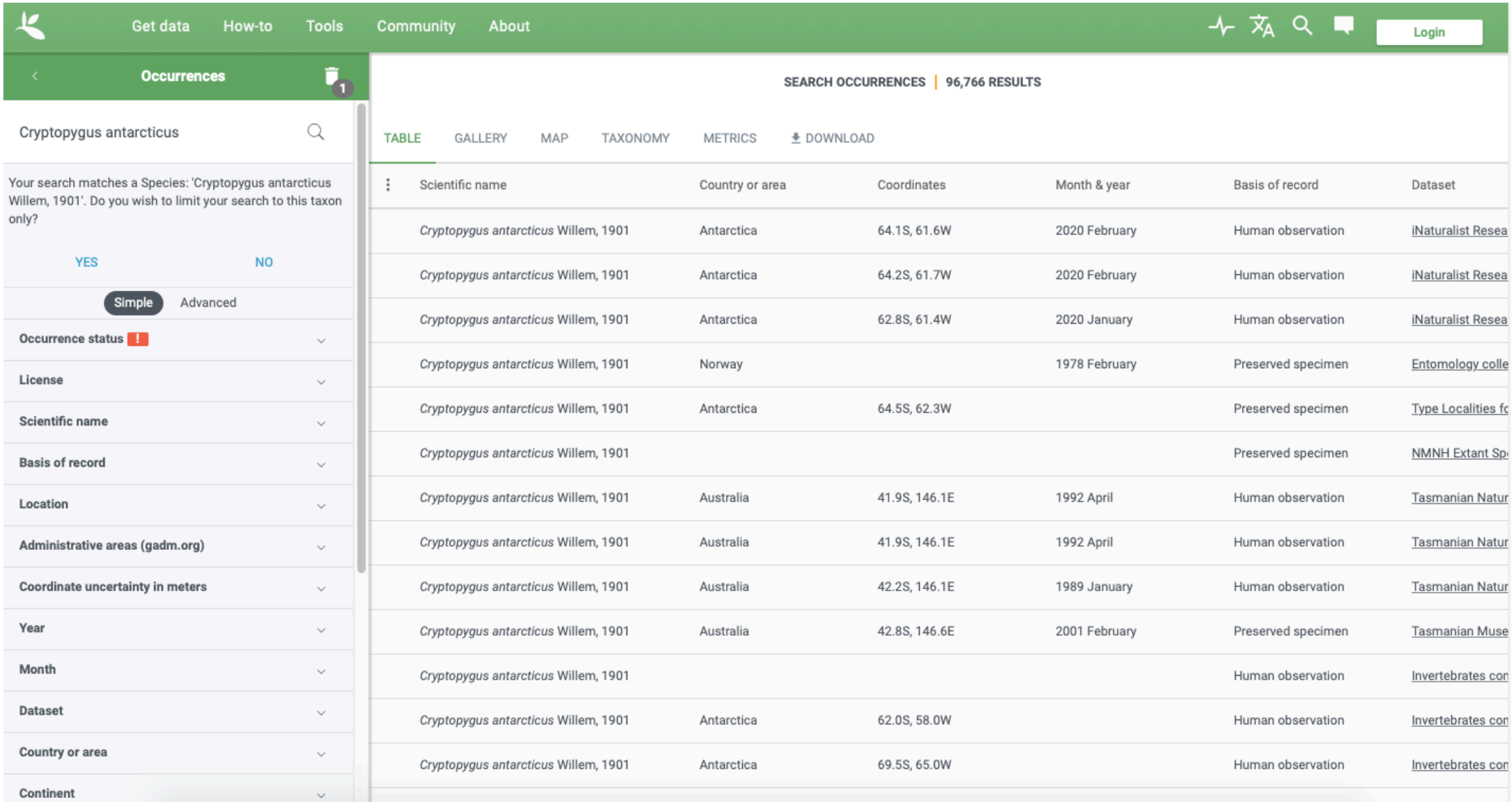
DEMO

Case study: Cryptopygus antarcticus



DEMO

Case study: Cryptopygus antarcticus



Scientific name	Country or area	Coordinates	Month & year	Basis of record	Dataset
Cryptopygus antarcticus Willem, 1901	Antarctica	64.1S, 61.6W	2020 February	Human observation	iNaturalist Resea
Cryptopygus antarcticus Willem, 1901	Antarctica	64.2S, 61.7W	2020 February	Human observation	iNaturalist Resea
Cryptopygus antarcticus Willem, 1901	Antarctica	62.8S, 61.4W	2020 January	Human observation	iNaturalist Resea
Cryptopygus antarcticus Willem, 1901	Norway		1978 February	Preserved specimen	Entomology colle
Cryptopygus antarcticus Willem, 1901	Antarctica	64.5S, 62.3W		Preserved specimen	Type Localities fc
Cryptopygus antarcticus Willem, 1901				Preserved specimen	NMNH Extant Sp
Cryptopygus antarcticus Willem, 1901	Australia	41.9S, 146.1E	1992 April	Human observation	Tasmanian Natur
Cryptopygus antarcticus Willem, 1901	Australia	41.9S, 146.1E	1992 April	Human observation	Tasmanian Natur
Cryptopygus antarcticus Willem, 1901	Australia	42.2S, 146.1E	1989 January	Human observation	Tasmanian Natur
Cryptopygus antarcticus Willem, 1901	Australia	42.8S, 146.6E	2001 February	Preserved specimen	Tasmanian Muse
Cryptopygus antarcticus Willem, 1901				Human observation	Invertebrates cor
Cryptopygus antarcticus Willem, 1901	Antarctica	62.0S, 58.0W		Human observation	Invertebrates cor
Cryptopygus antarcticus Willem, 1901	Antarctica	69.5S, 65.0W		Human observation	Invertebrates cor

Case study: Cryptopygus antarcticus

Get data | How-to | Tools | Community | About | Login

Cryptopygus antarcticus Willem, 1901	Antarctica	64.2S, 61.7W	2020 February	Human observation	iNat
Cryptopygus antarcticus Willem, 1901	Antarctica	62.8S, 61.4W	2020 January	Human observation	iNat
Cryptopygus antarcticus Willem, 1901	Norway		1978 February	Preserved specimen	Ento
Cryptopygus antarcticus Willem, 1901	Antarctica	64.5S, 62.3W		Preserved specimen	Type
Cryptopygus antarcticus Willem, 1901				Preserved specimen	NMN
Cryptopygus antarcticus Willem, 1901	Australia	41.9S, 146.1E	1992 April	Human observation	Tasr
Cryptopygus antarcticus Willem, 1901	Australia	41.9S, 146.1E	1992 April	Human observation	Tasr
Cryptopygus antarcticus Willem, 1901	Australia	42.2S, 146.1E	1989 January	Human observation	Tasr
Cryptopygus antarcticus Willem, 1901	Australia	42.8S, 146.6E	2001 February	Preserved specimen	Tasr

Missing information

Dataset

Country or area

Continent

Cryptopygus antarcticus Willem, 1901	Antarctica	62.0S, 58.0W	Human observation	Invertebrates con
Cryptopygus antarcticus Willem, 1901	Antarctica	69.5S, 65.0W	Human observation	Invertebrates con

Case study: Cryptopygus antarcticus

Get data | How-to | Tools | Community | About | Login

Cryptopygus antarcticus Willem, 1901	Antarctica	64.2S, 61.7W	2020 February	Human observation	iNat
Cryptopygus antarcticus Willem, 1901	Antarctica	62.8S, 61.4W	2020 January	Human observation	iNat
Cryptopygus antarcticus Willem, 1901	Norway		1978 February	Preserved specimen	Ento
Cryptopygus antarcticus Willem, 1901	Antarctica	64.5S, 62.3W		Preserved specimen	Type
Cryptopygus antarcticus Willem, 1901				Preserved specimen	NMN
Cryptopygus antarcticus Willem, 1901	Australia	41.9S, 146.1E	1992 April	Human observation	Tasm
Cryptopygus antarcticus Willem, 1901	Australia	41.9S, 146.1E	1992 April	Human observation	Tasm
Cryptopygus antarcticus Willem, 1901	Australia	42.2S, 146.1E	1989 January	Human observation	Tasm
Cryptopygus antarcticus Willem, 1901	Australia	42.8S, 146.6E	2001 February	Preserved specimen	Tasm

Misinterpretation of data standard

Missing information

Don’t forget down-stream cleaning of the data before use!!

Dataset

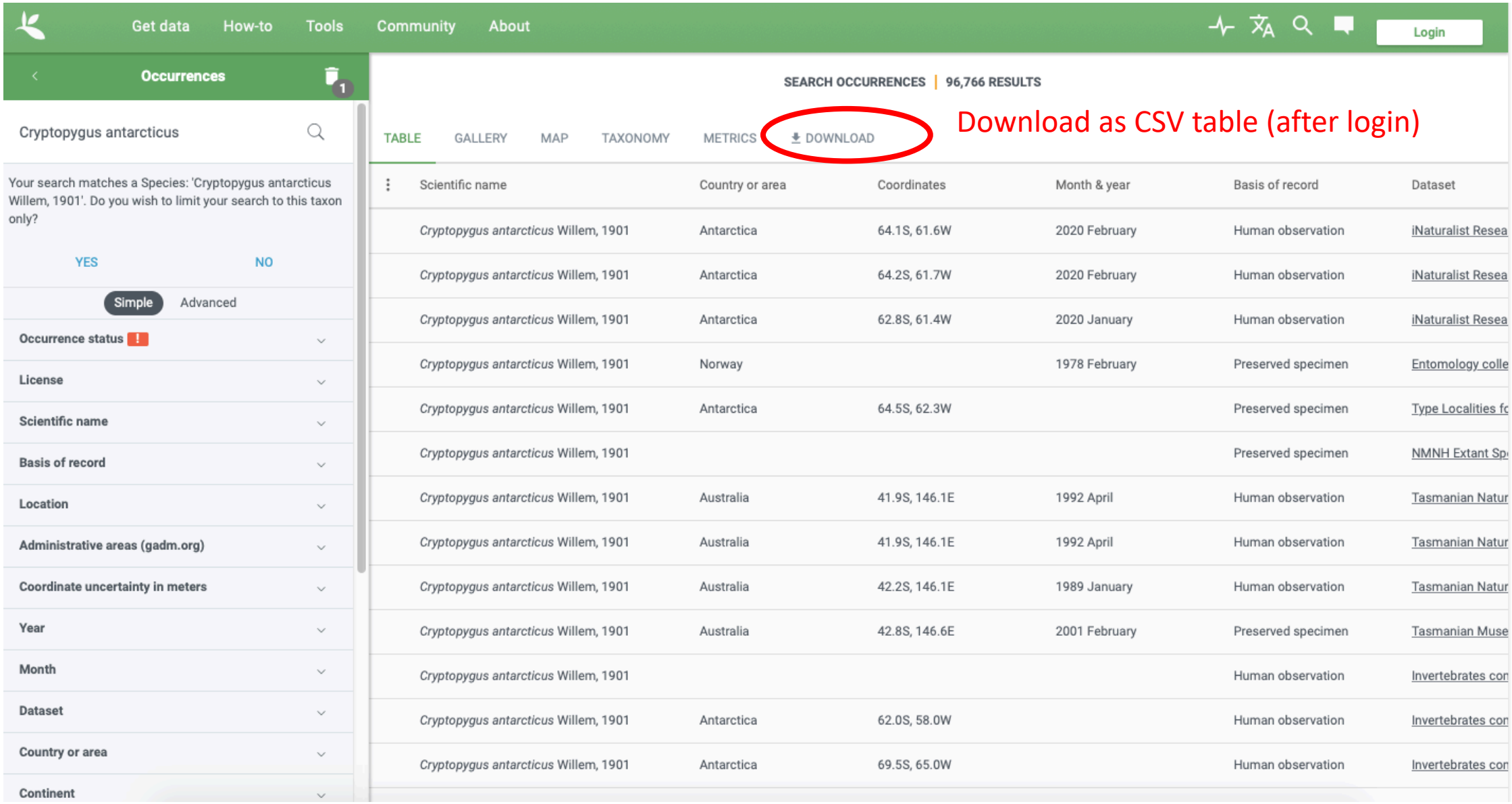
Country or area

Continent

			Human observation	Invertebrates con
Cryptopygus antarcticus Willem, 1901	Antarctica	69.5S, 65.0W	Human observation	Invertebrates con

DEMO

Case study: Cryptopygus antarcticus



DEMO

Case study: Cryptopygus antarcticus

Text Search

Uses EBI Search to perform a free text search across ENA data. For more detailed usage please refer to the help & documentation section.

Search term: Cryptopygus antarcticus — Search

Search results for Cryptopygus antarcticus

- **Sequence**
  - Sequence (487)
  - Sequence (Standard) (487)
- **Coding**
  - Coding (Update) (423)
  - Coding (Release) (423)
- **Non-coding**
  - Non-coding (107)
  - Non-coding (Standard) (107)
- **Taxon**
  - Taxon (21)

Sequence View all 487 results.

FJ648736	Cryptopygus antarcticus endo-beta-1,4-glucanase gene, complete cds.

Sequence (Standard) View all 487 results.

FJ648736	Cryptopygus antarcticus endo-beta-1,4-glucanase gene, complete cds.

Coding (Update) View all 423 results.

ACV50414	Cryptopygus antarcticus endo-beta-1,4-glucanase

Coding (Release) View all 423 results.

ACV50414	Cryptopygus antarcticus endo-beta-1,4-glucanase

Non-coding View all 107 results.

MK433191.1:169..237:tRNA	Cryptopygus antarcticus travei tRNA-Met

Non-coding (Standard) View all 107 results.

MK433191.1:169..237:tRNA	Cryptopygus antarcticus travei tRNA-Met

Taxon View all 21 results.

Case study: Cryptopygus antarcticus

Sequence (Standard) **Download ENA records:** FASTA TEXT

Accession	Description
FJ648736	Cryptopygus antarcticus endo-beta-1,4-glucanase gene, complete cds.
FJ648735	Cryptopygus antarcticus endo-beta-1,4-glucanase mRNA, complete cds.
EU021047	Cryptopygus antarcticus beta-1,4-mannanase precursor, mRNA, complete cds.
EU559744	Cryptopygus antarcticus beta-1,3-D-glucanase precursor, gene, complete cds.
FJ648734	Cryptopygus antarcticus endo-beta-1,3-glucanase precursor, mRNA, complete cds.
MK433191	Cryptopygus antarcticus travei mitochondrion, complete genome.
EU016194	Cryptopygus antarcticus mitochondrion, complete genome.
HM212750	Cryptopygus antarcticus antarcticus haplotype Caantarcticus_PEN1_2 28S ribosomal RNA gene, partial sequence.

Case study: Cryptopygus antarcticus

Sequence (Standard) **Download ENA records:** FASTA TEXT

Accession	Description
FJ648736	Cryptopygus antarcticus endo-beta-1,4-glucanase gene, complete cds.
FJ648735	Cryptopygus antarcticus endo-beta-1,4-glucanase mRNA, complete cds.
EU021047	Cryptopygus antarcticus beta-1,4-mannanase precursor, mRNA, complete cds.
EU559744	Cryptopygus antarcticus beta-1,3-D-glucanase precursor, gene, complete cds.
FJ648734	Cryptopygus antarcticus endo-beta-1,3-glucanase precursor, mRNA, complete cds.
MK433191	Cryptopygus antarcticus travei mitochondrion, complete genome.
EU016194	Cryptopygus antarcticus mitochondrion, complete genome.
HM212750	Cryptopygus antarcticus antarcticus haplotype Caantarcticus_PEN1_2 28S ribosomal RNA gene, partial sequence.

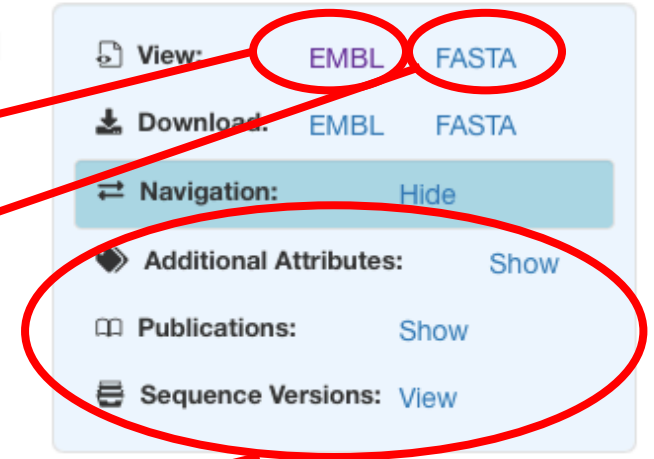
Case study: Cryptopygus antarcticus

Sequence: MK433191.1

Cryptopygus antarcticus travei mitochondrion, complete genome.

Organism:	Cryptopygus antarcticus travei
Accession:	MK433191
Mol Type:	genomic DNA
Topology:	CIRCULAR
Base Count:	15743

Show More



The metadata as XML text

The sequence (no metadata)

View metadata (="attributes") in browser

Navigation & Cross References

- Taxon:
  Taxon:359048

```
ID   MK433191; SV 1; circular; genomic DNA; STD; INV; 15743 BP.
XX
AC   MK433191;
XX
DT   07-NOV-2019 (Rel. 142, Created)
DT   07-NOV-2019 (Rel. 142, Last updated, Version 1)
XX
DE   Cryptopygus antarcticus travei mitochondrion, complete genome.
XX
KW   .
XX
OS   Cryptopygus antarcticus travei
OC   Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Collembola;
OC   Entomobryomorpha; Isotomoidea; Isotomidae; Anurophorinae; Cryptopygus;
OC   Cryptopygus antarcticus complex.
OG   Mitochondrion
XX
RN   [1]
RP   1-15743
RA   Monsanto D.M., Jansen van Vuuren B., Jagatap H., Jooste C.M.,
RA   Janion-Scheepers C., Teske P.R., Emami-Khoyi A.;
RT   "The complete mitogenome of the springtail Cryptopygus antarcticus travei
RT   provides evidence for speciation in the Sub-Antarctic region";
RL   Mitochondrial DNA B Resour 4(1):1195-1197(2019).
XX
RN   [2]
RP   1-15743
RA   Monsanto D.M., van Vuuren B., Jagatap H., Jooste C.M., Janion-Scheepers C.,
RA   Teske P.R., Emami-Khoyi A.;
RT   ;
RL   Submitted (22-JAN-2019) to the INSDC.
RL   Zoology, University of Johannesburg, University road, Johannesburg 2006,
RL   South Africa
XX
DR   MD5; c27daabe23edd33996a57d8b169e9a80.
XX
CC   ##Assembly-Data-START##
CC   Assembly Method       :: NOVOPlasty v. 2.7.2
CC   Assembly Name         :: Cryptopygus antarcticus travei
CC   Sequencing Technology :: Illumina
CC   ##Assembly-Data-END##
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..15743
FT                   /organism="Cryptopygus antarcticus travei"
FT                   /mol_type="mitochondrion"
FT                   /isolation_source="travei"
```

Case study: Antarctic peninsula amplicon data

Text Search

Uses EBI Search to perform a free text search across ENA data. For more detailed usage please refer to the help & documentation section.

Search term: antarctic peninsula amplicon — Search

Search results for antarctic peninsula amplicon

- **Read**
  - Experiment (784)
- **Study**
  - Study (7)
  - Study (Sequence) (7)
- **Sample**
  - Sample (48)
- **About**
  - ENA (2)

Experiment View all 784 results.

SRX525897	Ion Torrent PGM sequencing; Post-light based sequencing technology reveals distribution and interaction patterns of bacterial communities in an ornithogenic soil profile of Seymour Island, Antarctic Peninsula

Study View all 7 results.

ERP010485	Spitsbergen Island and Antarctic Peninsula Metagenomic Analysis

Study (Sequence) View all 7 results.

PRJEB37594	Characterization of microbial communities in initial soils of King George Island, Antarctic Peninsula region

Sample View all 48 results.

SAMN12015836	Amplicon of Seawater O'Higgins 2

ENA View all 2 results.

individual samples

Complete studies

Case study: Antarctic peninsula amplicon data

Secondary Study Accession: SRP114797

Study Title: Soil crust metagenome Antarctica

Center Name: Universidad de Chile

Study Name: soil crust metagenome

XML of the metadata

View: XML
XML (STUDY)
Download: XML
XML (STUDY)
Navigation: Show
Read Files: Hide
Additional Attributes: Show
Publications: Show
Related ENA Records: Show

Read Files

Show Column Selection

Download report: JSON TSV

Download Files as ZIP

Download selected files

Download All

Study Accession	Sample Accession	Experiment Accession	Run Accession	Tax Id	Scientific Name	FASTQ FTP	Subm
PRJNA397058	SAMN07443940	SRX3059861	SRR5894242	410658	soil metagenome	SRR5894242.fastq.gz	
PRJNA397058	SAMN07443903	SRX3059860	SRR5894243	410658	soil metagenome	SRR5894243.fastq.gz	
PRJNA397058	SAMN07443902	SRX3059859	SRR5894244	410658	soil metagenome	SRR5894244.fastq.gz	
PRJNA397058	SAMN07443905	SRX3059858	SRR5894245	410658	soil metagenome	SRR5894245.fastq.gz	
PRJNA397058	SAMN07443904	SRX3059857	SRR5894246	410658	soil metagenome	SRR5894246.fastq.gz	

metadata

To download data