What about metadata?

Tovo Rabemanantsoa

To cite this version:

Tovo Rabemanantsoa. What about metadata?. 2024. hal-04645703

HAL Id: hal-04645703

https://hal.inrae.fr/hal-04645703v1

Submitted on 12 Jul 2024

What about the metadata?

REacTION 2024

What is metadata?

Definition

Data that provides information about other data

It can describe a collection, a single resource, or a component part of a larger resource

What is metadata?

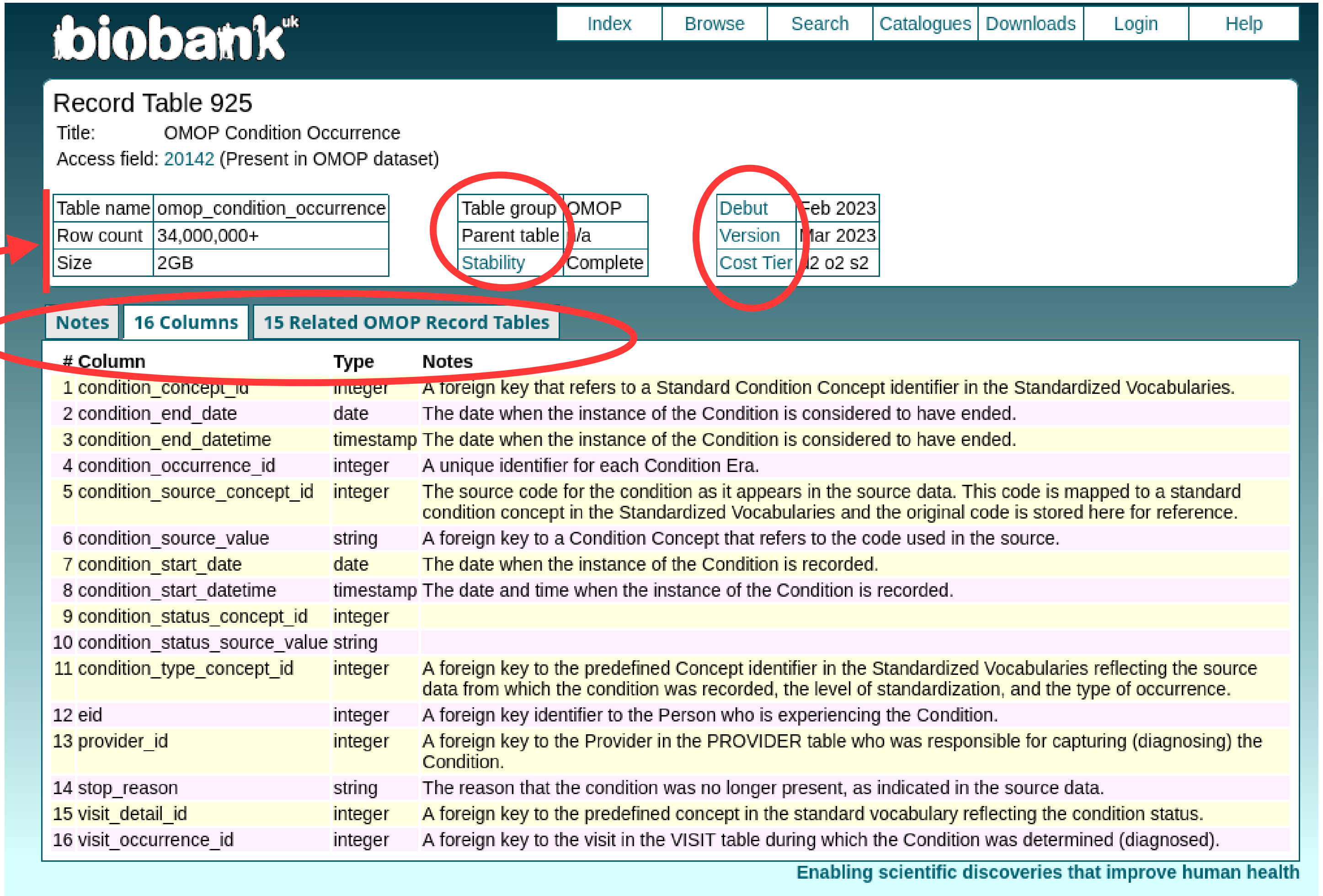
Types of metadata

- Descriptive (title, abstract, author, keywords,...)
- Structural (how compound objects are put together)
- Administrative (how it was created, file type, access rights,...)

What is metadata?

How does it look like?

All of these are metadata



Record Table 925

Title: OMOP Condition Occurrence

Access field: 20142 (Present in OMOP dataset)

Table name	omop_condition_occurrence
Row count	34,000,000+
Size	2GB

Table group	OMOP
Parent table	n/a
Stability	Complete

Debut	Feb 2023
Version	Mar 2023
Cost Tier	d2 o2 s2

#	Column	Type	Notes
1	condition_concept_id	integer	A foreign key that refers to a Standard Condition Concept identifier in the Standardized Vocabularies.
2	condition_end_date	date	The date when the instance of the Condition is considered to have ended.
3	condition_end_datetime	timestamp	The date when the instance of the Condition is considered to have ended.
4	condition_occurrence_id	integer	A unique identifier for each Condition Era.
5	condition_source_concept_id	integer	The source code for the condition as it appears in the source data. This code is mapped to a standard condition concept in the Standardized Vocabularies and the original code is stored here for reference.
6	condition_source_value	string	A foreign key to a Condition Concept that refers to the code used in the source.
7	condition_start_date	date	The date when the instance of the Condition is recorded.
8	condition_start_datetime	timestamp	The date and time when the instance of the Condition is recorded.
9	condition_status_concept_id	integer	
10	condition_status_source_value	string	
11	condition_type_concept_id	integer	A foreign key to the predefined Concept identifier in the Standardized Vocabularies reflecting the source data from which the condition was recorded, the level of standardization, and the type of occurrence.
12	eid	integer	A foreign key identifier to the Person who is experiencing the Condition.
13	provider_id	integer	A foreign key to the Provider in the PROVIDER table who was responsible for capturing (diagnosing) the Condition.
14	stop_reason	string	The reason that the condition was no longer present, as indicated in the source data.
15	visit_detail_id	integer	A foreign key to the predefined concept in the standard vocabulary reflecting the condition status.
16	visit_occurrence_id	integer	A foreign key to the visit in the VISIT table during which the Condition was determined (diagnosed).

Enabling scientific discoveries that improve human health

What is metadata?

How does it look like?

README content

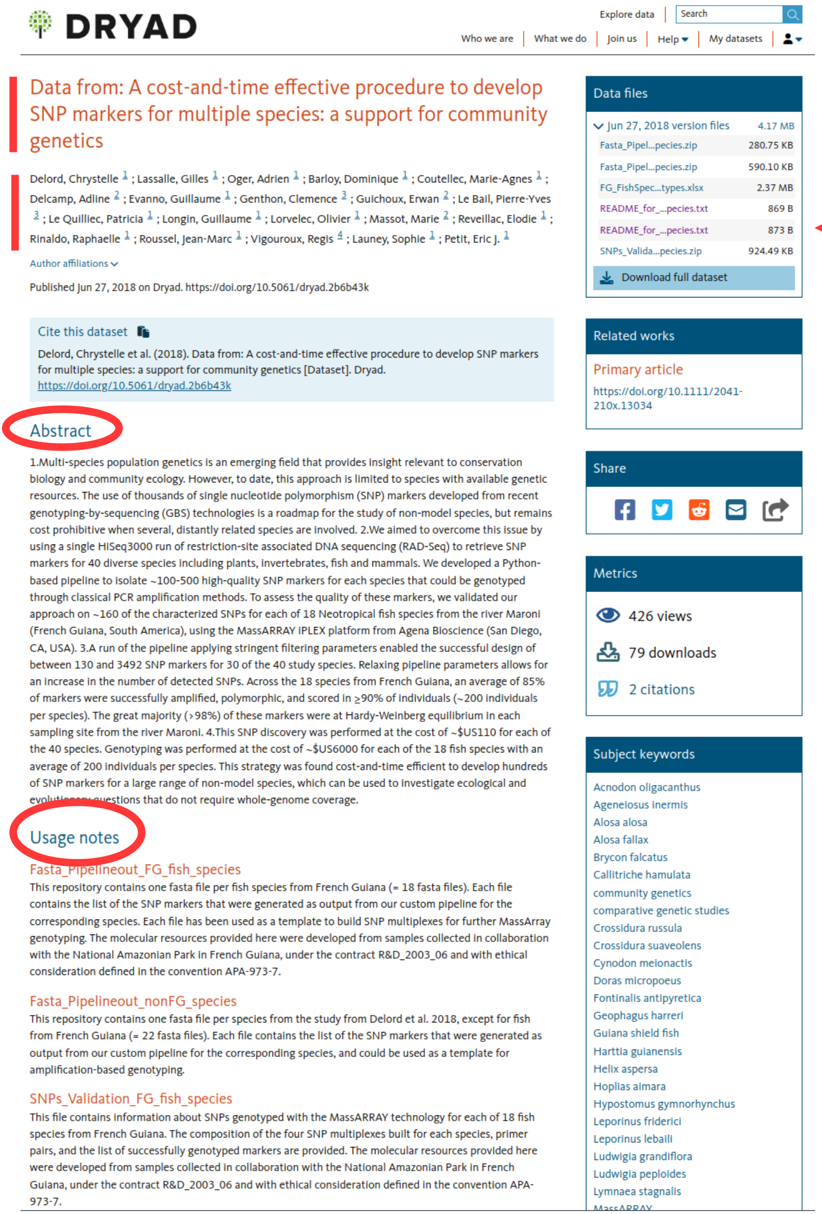
Following fasta sequences read as follows:

(Example from Acnodon oligacanthus)

>1302|69|[C/A]|AOL

```
TGCAGGTTATTTAACTCTTTTTCATAATGCACCTCATACATAGTAACCATTTTCAAACATGTCTTTC
ACCACCTGCGACGTGATATATTCAAAGTCTTATAGTTACAGTACATACTTTTTGTAGTTAAAATGG
AAAGAAGTGGCA
```

-The number '1302' is the identifiant of the sequence corresponding RAD-tag within the 'catalog_index' table of the Stacks database. This basically serves as marker ID.
-The number '69' is the position of the expected SNP marker. Please note that Stacks numerotation starts from 0; so a SNP at position '69' actually corresponds to the 70th nucleotide within the 145 bp sequence.
-Information within brackets [C/A] correspond to the two expected alleles for the SNP.
-'AOL' is a descriptive code relative to the species 'A'cnodon 'OL'igacanthus (First letter of genus, first and second letter of species)



title

authors

README

Data from: A cost-and-time effective procedure to develop SNP markers for multiple species: a support for community genetics

Delord, Chrystelle [1] ; Lassalle, Gilles [1] ; Oger, Adrien [1] ; Barloy, Dominique [1] ; Coutellec, Marie-Agnes [1] ; Delcamp, Adline [2] ; Evanno, Guillaume [1] ; Genthon, Clemence [3] ; Guichoux, Erwan [2] ; Le Bail, Pierre-Yves [3] ; Le Quilliec, Patricia [1] ; Longin, Guillaume [1] ; Lorvelec, Olivier [1] ; Massot, Marie [2] ; Reveillac, Elodie [1] ; Rinaldo, Raphaelle [1] ; Roussel, Jean-Marc [1] ; Vigouroux, Regis [4] ; Launey, Sophie [1] ; Petit, Eric J. [1]

Author affiliations

Published Jun 27, 2018 on Dryad. https://doi.org/10.5061/dryad.2b6b43k

Cite this dataset

Delord, Chrystelle et al. (2018). Data from: A cost-and-time effective procedure to develop SNP markers for multiple species: a support for community genetics [Dataset]. Dryad. https://doi.org/10.5061/dryad.2b6b43k

Abstract

1.Multi-species population genetics is an emerging field that provides insight relevant to conservation biology and community ecology. However, to date, this approach is limited to species with available genetic resources. The use of thousands of single nucleotide polymorphism (SNP) markers developed from recent genotyping-by-sequencing (GBS) technologies is a roadmap for the study of non-model species, but remains cost prohibitive when several, distantly related species are involved. 2.We aimed to overcome this issue by using a single HiSeq3000 run of restriction-site associated DNA sequencing (RAD-Seq) to retrieve SNP markers for 40 diverse species including plants, invertebrates, fish and mammals. We developed a Python-based pipeline to isolate ~100-500 high-quality SNP markers for each species that could be genotyped through classical PCR amplification methods. To assess the quality of these markers, we validated our approach on ~160 of the characterized SNPs for each of 18 Neotropical fish species from the river Maroni (French Guiana, South America), using the MassARRAY iPLEX platform from Agena Bioscience (San Diego, CA, USA). 3.A run of the pipeline applying stringent filtering parameters enabled the successful design of between 130 and 3492 SNP markers for 30 of the 40 study species. Relaxing pipeline parameters allows for an increase in the number of detected SNPs. Across the 18 species from French Guiana, an average of 85% of markers were successfully amplified, polymorphic, and scored in ≥90% of individuals (~200 individuals per species). The great majority (>98%) of these markers were at Hardy-Weinberg equilibrium in each sampling site from the river Maroni. 4.This SNP discovery was performed at the cost of ~$US110 for each of the 40 species. Genotyping was performed at the cost of ~$US6000 for each of the 18 fish species with an average of 200 individuals per species. This strategy was found cost-and-time efficient to develop hundreds of SNP markers for a large range of non-model species, which can be used to investigate ecological and evolutionary questions that do not require whole-genome coverage.

Usage notes

Fasta_Pipelineout_FG_fish_species

This repository contains one fasta file per fish species from French Guiana (= 18 fasta files). Each file contains the list of the SNP markers that were generated as output from our custom pipeline for the corresponding species. Each file has been used as a template to build SNP multiplexes for further MassArray genotyping. The molecular resources provided here were developed from samples collected in collaboration with the National Amazonian Park in French Guiana, under the contract R&D_2003_06 and with ethical consideration defined in the convention APA-973-7.

Fasta_Pipelineout_nonFG_species

This repository contains one fasta file per species from the study from Delord et al. 2018, except for fish from French Guiana (= 22 fasta files). Each file contains the list of the SNP markers that were generated as output from our custom pipeline for the corresponding species, and could be used as a template for amplification-based genotyping.

SNPs_Validation_FG_fish_species

This file contains information about SNPs genotyped with the MassARRAY technology for each of 18 fish species from French Guiana. The composition of the four SNP multiplexes built for each species, primer pairs, and the list of successfully genotyped markers are provided. The molecular resources provided here were developed from samples collected in collaboration with the National Amazonian Park in French Guiana, under the contract R&D_2003_06 and with ethical consideration defined in the convention APA-973-7.

Data files

Jun 27, 2018 version files	4.17 MB
Fasta_Pipel...pecies.zip	280.75 KB
Fasta_Pipel...pecies.zip	590.10 KB
FG_FishSpec...types.xlsx	2.37 MB
README_for_...pecies.txt	869 B
README_for_...pecies.txt	873 B
SNPs_Valida...pecies.zip	924.49 KB

Download full dataset

Related works

Primary article

https://doi.org/10.1111/2041-210x.13034

Share

Metrics

426 views

79 downloads

2 citations

Subject keywords

Acnodon oligacanthus
Ageneiosus inermis
Alosa alosa
Alosa fallax
Brycon falcatus
Callitriche hamulata
community genetics
comparative genetic studies
Crossidura russula
Crossidura suaveolens
Cynodon meionactis
Doras micropoeus
Fontinalis antipyretica
Geophagus harreri
Guiana shield fish
Harttia guianensis
Helix aspersa
Hoplias aimara
Hypostomus gymnorhynchus
Leporinus friderici
Leporinus lebaili
Ludwigia grandiflora
Ludwigia peploides
Lymnaea stagnalis

Why is it matters?

Role of the metadata

At least, gives an answer to the:

- What?
- How?
- Who?

Provides:

- Differentiation between datasets
- Context (scope, units,...)
- Reassurance of reliability

→ a keystone of FAIR data

Why is it matters?

Reproducibility and reusability

- Metadata enforces the FAIRness of your data
- FAIRness is a major step to reproducibility and reusability
- Reproducibility and reusability leads to a stronger science

Why is it matters?

It’s not (only) about the money

Not having or having poor quality metadata

- Is a waste of time (to find/retrieve, to understand/authenticate/analyse)
- Is a waste of resources (redundant copies, collection of existing data)

→ It costs money (€10bn/y in Europe)

Some good practices

Scaffold a strategy

- Define a data governance strategy
- Build a cross-functional data team
- Adopt a standard
- Deploy a metadata management tool

Some good practices

Stick to the plan

- Regularly assess your alignment with your strategy
- Make sure that your teammates are:
  - Up to date on their skills
  - Have the resources to implement the strategy
  - Still ok with the strategy

Some good practices

Stay agile

If the strategy no longer suits you (or you team):

- Change it!
- Work out the change with your data team
- Communicate vastly

Standards

Datacite Metadata Schema: a list of core metadata properties chosen for an accurate and consistent identification of a resource for citation and retrieval purposes, along with recommended use instructions - https://schema.datacite.org/

Dublin Core: A basic, domain-agnostic standard which can be easily understood and implemented, and as such is one of the best known and most widely used metadata standards - http://dublincore.org/

RO-Crate: RO-Crate is a community effort to establish a lightweight approach to packaging research data with their metadata - https://www.researchobject.org/ro-crate/

And tools

ODAM: Open Data for Access and Mining is an Experimental Data Table Management System (EDTMS) - https://inrae.github.io/ODAM/

MOLGENIS: a data platform for researchers to accelerate scientific collaborations
and for bioinformaticians who want to make researchers happy - https://molgenis.org/

PDBx/mmCIF: a standard Protein Data Bank archive format - https://mmcif.wwpdb.org/

Dspace: The DSpace digital asset management system that powers your Institutional Repository - https://wiki.lyrasis.org/display/DSDOC8x/

DRYAD: Open data publishing platform and a community committed to the open availability
and routine re-use of all research data - https://datadryad.org

FAIR Aware: FAIR-Aware helps you assess your knowledge of the FAIR Principles,
and better understand how making your data(set) FAIR can increase the potential value and impact of your data -
https://fairaware.dans.knaw.nl/

Recherche Data gouv: Un écosystème au service du partage et de l'ouverture des données de la recherche - https://recherche.data.gouv.fr

References

- PwC EU Services (2018). Cost of not having FAIR research data - Cost-Benefit analysis for FAIR research data. DOI: 10.2777/02999
- NISO Press (2004). Understanding Metadata. ISBN: 1-880124-62-9
- Anne Marie Smith, Ph.D. (2017). Turning Data into Knowledge: Creating and Implementing a Meta Data Strategy
- Pasquetto, I V et al 2017 On the Reuse of Scientific Data. Data Science Journal, 16: 8, pp. 1–9, DOI: 10.5334/dsj-2017-008
- Barriocanal, Elena & Sánchez-Alonso, Salvador & Sicilia, M.. (2017). Deploying Metadata on Blockchain Technologies. 38-49. DOI: 10.1007/978-3-319-70863-8_4
- Iturbide, M., Fernández, J., Gutiérrez, J.M. et al. Implementation of FAIR principles in the IPCC: the WGI AR6 Atlas repository. Sci Data 9, 629 (2022). https://doi.org/10.1038/s41597-022-01739-y
- Metadata Standards Catalog: https://rdamsc.bath.ac.uk/