ARPHA Preprints

PENSOFT.

Software Description

Author-formatted document posted on 30/09/2024

Published in a RIO article collection by decision of the collection editors.

DOI: https://doi.org/10.3897/arphapreprints.e138026

Developing the ParAqua database: methodology and implications

Andrea Tarallo, Giuseppe Turrisi, Davide Raho, Ilaria Rosati

Developing the ParAqua database: methodology and implications

Andrea Tarallo[1*], Giuseppe Turrisi[2], Davide Raho[1], Ilaria Rosati[1]

[1]Institute of Research on Terrestrial Ecosystems (IRET), National Research Council (CNR), Lecce, Italy

[2]LifeWatch ERIC, Service Centre, S.P. Lecce-Monteroni – Ecotekne, Lecce, 73100, Italy
[*]Corresponding author: andrea.tarallo@cnr.it

Abstract— This paper presents the collaborative efforts of Working Group 1 (WG1 - Occurrence and detection of zoosporic parasites) and Working Group 2 (WG2 - Drivers underlying the dynamic of zoosporic diseases) within the ParAqua COST Action, a research initiative focused on understanding zoosporic parasites and their interactions with algae. Initially conceived as an interactive web page, one of the deliverables of WG1 has evolved into a centralised database for data gathered by the scientific community of ParAqua. In this paper we present a summary of our work, carried out from July 2022 to October 2023. After gathering and analysing community needs, we have harmonised data collection, designed and implemented the database structure. The upcoming steps involve the data upload and integration process into the database and creating a Graphical User Interface to query the database. The paper focuses on the methodology and discusses the challenges faced during the activity. As part of our commitment to promoting the practices of open science, we wrote this paper to document the entire process of the database development. Our aim is to provide a clear pathway for others to expand, challenge, and refine our database designing process, encouraging a dynamic exchange of ideas that propels the field forward. In a broad sense, this research contributes to the understanding of algae-parasite interactions, by providing a unique experience of data mobilisation and integration in the field of algae parasites research

Keywords: Zoosporic parasite — algae — parasite-algae interaction — COST Action

1. Introduction

In the intricate tapestry of natural ecosystems, algae stand as fundamental contributors, fuelling the planet's oxygen supply and supporting the aquatic food web (Naselli-Flores & Padisák 2023). Like any other living organism, however, algae are vulnerable to parasitic infections (Grami et al., 2011).

Through detailed exploration of the interactions between parasites and their alga hosts, researchers are elucidating crucial insights concerning ecological stability, biodiversity, and the aquatic ecological network. The study of algal parasites is not just limited to the academic interest. Understanding algae-parasite interactions provides fundamental insights essential for sustainable resource management and biotechnology (Carney & Lane, 2014).

The COST Action ParAqua (Rasconi et al., 2022), funded through the agency for research and innovation networks COST (www.cost.eu), proposes an innovative European network connecting scientists, industries, and water-body managers to address a critical gap in our understanding of zoosporic parasites, fungi, and fungi-like aquatic microorganisms that impact algae natural populations and have substantial economic implications in the microalgae biotech industry.

To expand and integrate the existing knowledge, ParAqua aims to compile and disseminate comprehensive information on zoosporic parasites' occurrence and their relationship with hosts. This aim is exemplified by the work that would have been carried out by Working Group 1 (WG1) and Working Group 2 (WG2).

In particular, WG1 focuses on collecting data regarding the occurrence of zoosporic parasites in natural and industrial systems. Their efforts will culminate in the creation of an interactive webpage to display information on aquatic zoosporic parasites, their hosts, and their distribution. In parallel, WG2 aims to write review articles compiling literature data on the environmental drivers of zoosporic parasite infection. This includes exploring the potential for rapid evolution leading to shifts in host-parasite interactions. Additionally, they will create fact sheets explaining key environmental drivers, both abiotic and biotic, involved in zoosporic parasite infections.

Both working groups share a common strategy: collecting data from various sources and of different natures on zoosporic parasites, centralising this data in review articles, reports, or browsable web pages, and using them to conduct meta-analyses to discover potential underlying correlations between variables that can be applied in the management of parasites in both natural and artificial systems.

The shared strategy highlights the collaborative nature of their work, triggering the expansion of what was initially conceived as an "*interactive web page [...] to integrate knowledge and interactive sharing of* *information by the users*" (Rasconi et al., 2022). This has evolved into a more sophisticated tool: a centralised database (DB) to collect and share all the different data coming from WG1 and WG2 activities. In this endeavour, the Lecce unit of the Institute of Research on Terrestrial Ecosystems of the National Research Council of Italy (CNR-IRET-LE) was involved as partner in the Action to lead this activity. CNR-IRET-LE has extensive expertise in managing digital resources for environmental scientists and leads data management activities in LifeWatch Italy (https://www.lifewatchitaly.eu/en/home-english/), the Italian node of the European Research Infrastructure Consortium LifeWatch ERIC, an e-Science Infrastructure on biodiversity and ecosystem research. LifeWatch Italy focuses on data management, curation, harmonisation, and web services development.

The purpose of the DB is to systematically organise the data gathered during the various activities of ParAqua in a standardised and centralised digital entity to facilitate their access and reuse. Although there is no shared definition, this process is commonly referred to as data mobilisation (Diack et al., 2022, Baker et al., 2015; Schulman et al., 2021).

In the first instance, the DB is useful for the ParAqua community, allowing to conduct meta-analyses on a set of harmonised variables, thus achieving one of the objectives of the Action. Secondly, we foreseen the use of such tool beyond ParAqua, promoting the research and study of interactions between parasites and algae.

As part of our commitment to promoting the practices of open science, we wrote this paper to document the entire process of the DB development. This documentation describes the activities undertaken to design the ParAqua DB, outlines the strategies we adopted, and highlights some lessons learned during the process.

The activities are summarised in the following four steps:

i. Community needs gathering, DB requirements definition and harmonisation;
ii. DB conceptual design;
iii. DB logical design;
iv. DB physical design.

Our aim is to provide a clear pathway for others to expand, challenge, and refine our DB designing process, encouraging a dynamic exchange of ideas that propels the field forward.

2. Methodology

2.1 Community needs gathering, DB requirements definition and harmonisation

The first challenge was to collect data coming from different experiences and bring them together in a harmonised way. As it is usual in the data mobilisation activities that we carried out in LifeWatch Italy, we started with a preliminary survey to understand the type and volume of data that we should expect. This is also needed to outline a strategy to manage the data that we will receive. The survey usually contains a very minimal set of questions with prefilled answers to help and harmonise the collection of the information. A complete list is provided in Table 1.

The second step was the actual kick-off of the mobilisation activity. We did it with a hybrid workshop held in Cyprus on the 5th-6th of July 2022 (Rasconi 2023). During the workshop, we carried out a data management literacy workshop, during which we agreed on the basic concepts of the research data management cycle, and on the terminology, we would be using from the subsequent activities.

TABLE I

Column header	**Help text**	**Compulsory field**
ID	Consecutive numbers (in case you are describing more than one data resource). Please use one raw for each data resource.	No
Name	First name + family name	Yes
Email contact	Your primary email. It will be used to contact you back for the subsequent data gathering process	Yes
Affiliation	Your primary affiliation	Yes
Data resource title or short description	The title of the data resources, for instance the title of the paper that include the data resources, the database name, or, in case this information is missing, a short description of what is included in the data resources that will be provided	Yes
Is it already published? If yes please provide the access link	The persistent link to the published resource (e.g. URI, DOI, etc.). This can for instance be the link to the paper supplementary material in which your data are contained, the link to the repository where the data are stored, the link of a database, etc.	No
Data category	Choose from a dropdown menu if your data resources concern primarily the parasites, their hosts, or measured environmental variables. In case none of the previous descriptions match your data, choose "other".	Yes
Data type	Choose from a dropdown menu if your data are morphology/morphometric	

	measurements, occurrences such as species lists, sequence data, or environmental variables such as concentrations, chemico-physical parametres, etc.	
Artificial/Natural system	Choose from a dropdown menu in what kind of (natural or artificial) system your data resource is based	Yes
Habitat	Choose from a dropdown menu the type of habitat that better describes the provenance of your resource. The habitat classification is based on the first level of EUNIS classification.	Yes
Spatial range (local/country/subcontinental)	Select the spatial extent of your dataset: - Local: such as single site or sub-regional sampling; - Country: i.e. regional to whole country level; - Subcontinental: the sampling was performed over more than one country	No
Temporal range	Select the temporal extent of your dataset (e.g. "2003"; "from 2018 onward"; "every summer from 1998 to 2014")	No
Data format type	Select if the data resource file is in the format of a: - Tabular data (e.g. xlsx, csv, tsv; etc.) - Text (e.g. plain text, pdf, docx, html; etc:) - Images (e.g. TIFF, jpeg, png, etc.) - Structured data (e.g. database, relational database, XML, rdf, other)	Yes
Are metadata present?	The metadata is the information that is used to describe the data resources, often organized in standardized schema (e.g. the Ecological Metadata Language - EML). Metadata contain information about the people who collected and organized the data resource, who published the resource, the sampling protocol, the name of the sampling campaign, etc.	No
Dimension (Number of bytes)	The dimension of the file(s) in bytes	Yes
Notes	Any other comment you think are important for us to know	No

Central to the ParAqua DB design process was the gathering of requirements, identifying the data to be stored and the relationships between the data. The strategy to harmonise all the different data was to build, together with the scientific community of the project, a series of templates to be filled with the preexisting data. The requirements collection was initiated during the same workshop in Cyprus.

Three streams of data were initially defined (Figure 1):

v. Data coming from in situ observations.

This first stream of data is mainly coming from the ParAqua members, consisting in data about observation of algae and parasites presence in water bodies, along with other complementary variables, mainly environmental ones and ecomorphological traits of the recorded organisms. The activity was led by WG1. We supported the researchers in agreeing on the templates for the data gathering of *in situ* observations, e.g. collected as three types of information: the presence of the parasite along with the associated host, traits (characteristics of the parasite), and accessory environmental variables. This information is stored in spreadsheets[1].

We then drove the scientific community to use the templates using asynchronous meetings. A video was recorded on how to fill the templates. After having distributed it, several “Question & Answers” online meetings were organised during which we solved specific problems of harmonisation.

vi. Data coming from NCBI.

NCBI stands for the National Center for Biotechnology Information (Sayers et al., 2022). NCBI is a well-known and trusted source of data regarding molecular biology, bioinformatics, and genomics. All the genetic data (i.e. DNA sequences) related to parasites that have been published in the literature are virtually present in NCBI databases. Along with genetic data, other accompanying variables are usually recorded, such as the host from which the parasite has been isolated, the sampling location, etc. For this reason, another activity was led by WG1 to capture all those data. To do so, data related to some specific parasites are extracted and automatically downloaded with a customised query via a Python script. The data so downloaded were then manually curated and stored into customised spreadsheets[2].

vii. Literature data on environmental drivers.

[1] https://docs.google.com/spreadsheets/d/1Df4RinQF2GgS-r8Wor8Cy7uADWrY3rap/edit?usp=drive_web&ouid=113032675332595297709&rtpof=true

[2] https://docs.google.com/spreadsheets/d/10KdUR_95_iCvGSZsDnv9wy1f-Hgc8oeg/edit?rtpof=true#gid=1234904314

Carried out by WG2, this activity aims at collecting data from the literature about existing vectors in nature that favour parasite-host interaction extracted and saved in spreadsheets. E.g. how the pH changes with the change in parasite abundance, how the parasitic prevalence changes in relation to irradiation, etc. The activity is currently in development. A tentative template is available[3].

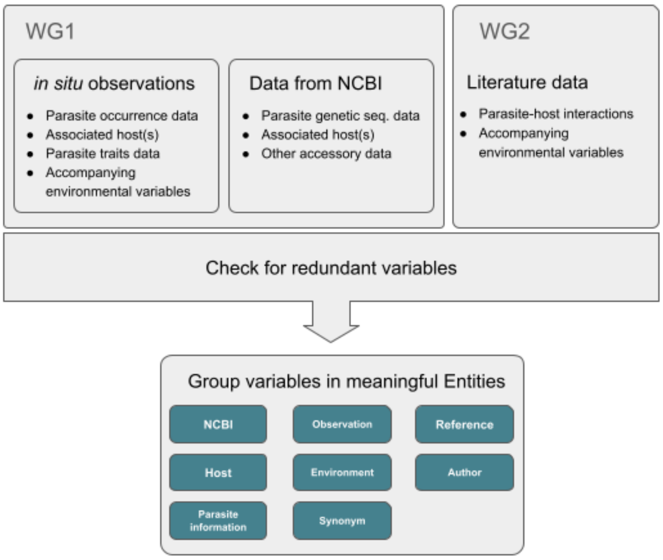

Fig. 1 Schematic description of the workflow we followed to collect the desired variables and group them into entities.

In the end, the number of variables that were foreseen to be collected using the templates described above was 108, without considering the full taxonomy of the parasites and host species (i.e. from kingdom to the species level), which may account for more than 30 additional variables. Since some of the considered variables were redundant, a first check to eliminate redundancies has been carried out. After that, the remaining unique variables were 66. When possible, the variables have been mapped with standard terminologies, as glossaries, vocabularies and thesauri. We primarily use Darwin Core terms (Darwin Core Maintenance Group, 2023) and the thesauri developed by LifeWatch Italy (https://ecoportal.lifewatch.eu/). The process of mapping is still in progress and the variables list showed in the current paper is still tentative and not completely matched with controlled terminologies. Once we have a list of unique standardised variables, we gathered them in “entities”, i.e. containers in which all the variables are conceptually related among them. For instance, in the entity “Parasite information” we gathered all the data that describe the parasite, such as its taxonomy and the traits that are only dependent by the species we are observing, like the presence or not of the flagella. All the variables that, instead, may change every time a parasite is observed, such as the temperature of the water during the sampling, the sampling coordinates, the density of the parasites, are gathered into the entity “Observation”. The process is depicted in Figure 2. We defined 8 entities:

i. Parasite information: the taxonomy of the parasite and all the fixed information about it (i.e. traits non-dependent on the life stage);
ii. Host: the taxonomy of the host algae and all the fixed information about it (i.e. traits non-dependent on the life stage);
iii. Observation: the variables that are depending on the sampling event;
iv. NCBI: the data that are collected from the NCBI DB;
v. Synonym: the synonym species of the parasites;
vi. Reference: the literature reference of the data that are gathered in the other entities;
vii. Author: the information related to the authors of the literature used.

The details are listed in the Appendix 1 tables.

[3] https://docs.google.com/spreadsheets/d/18ClrsQUZ6Sn3QSc4hNqEUDdThMDwMtCn/edit#gid=265040610

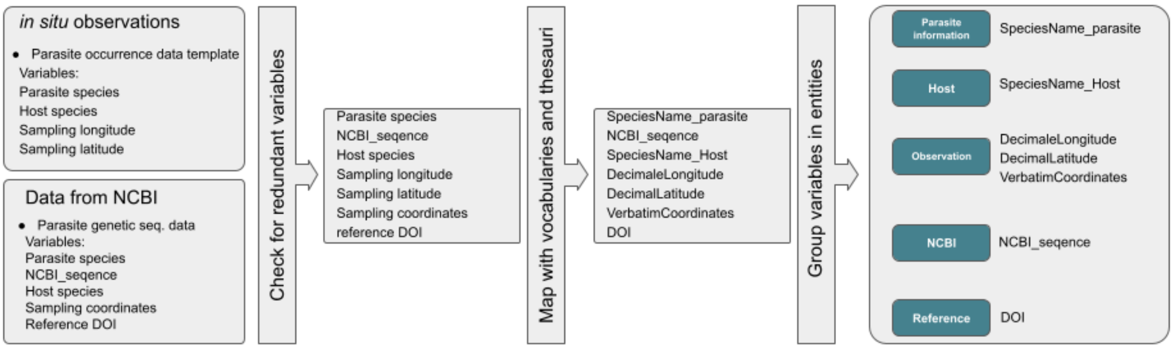

Fig. 2 The figure exemplifies the pipeline we applied to the variables collected by using the templates to obtain non-redundant and standardised variables grouped into entities.

2.1.1 Data Management Plan

In parallel, we developed the data management plan, which contains all the relevant information regarding how data is handled in the ParAqua Action. The structure of the data management plan follows the division into templates mentioned earlier: three templates for in-situ observations and one for data from NCBI. The collection of data on environmental drivers was not included in the current version, as it had not yet started when the data management plan was published. We used ARGOS (https://argos.openaire.eu/splash/), an online machine-actionable tool developed by OpenAIRE to facilitate the implementation of machine-actionable Data Management Plans. The first version of the ParAqua Data Management Plan was recently published (Tarallo et al., 2023).

2.2 Conceptual Design

The purpose of the conceptual design is to build a model that reflects the requirements and provides a clear understanding of the entities, their relationships and constraints. The goal at this stage is to design a DB that is independent of software and physical details. At this stage, we have identified the entities and their attributes (i.e. the variables). The practical objectives are to define the relationships between them, establish the constraints, and design a conceptual model called the Entity-Relationship model. The requirements collection was conducted through the analysis of raw data available in various formats and coming from the three sources listed above (*in situ* observations, NCBI, and literature). The entity-relationship diagram, which models how the entities relate to each other, is represented in Figure 3.

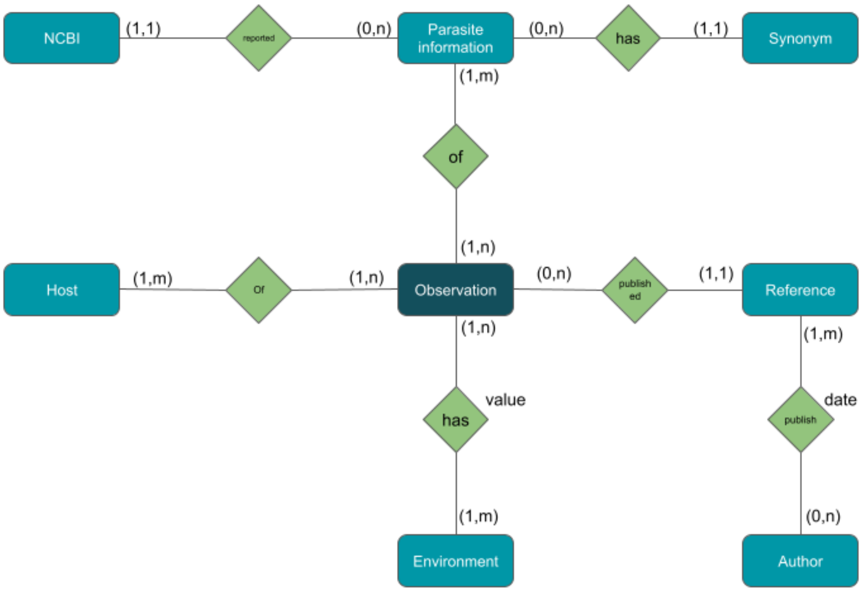

Fig. 3 Entity-Relationship model.

2.3 *Logical Design*

The logical DB design is the process of transforming the Entity-Relationship model into a relational schema. In other words, the Entity-Relationship model is extended to represent all the variables gathered within the entities (the tables), along with the primary and foreign keys. The purpose of keys is to establish a formal relationship between tables. This step is fundamental to allow the retrieving of information from the DB through queries. The result of this step is the relational model showed in Figure 4. The detail of each table content is available in the Appendix tables.

2.4 *Physical Design*

During the physical design process, the data gathered during the logical design phase was converted into a description of the physical DB structure. The DB management system (DBMS) on which the data model of the logic scheme will be implemented is MySql. In Figure 4 the relational schema is represented. All the Data Definition Language (DDL) commands needed for the creation of the tables in the DBMS are listed in the Appendix 2.

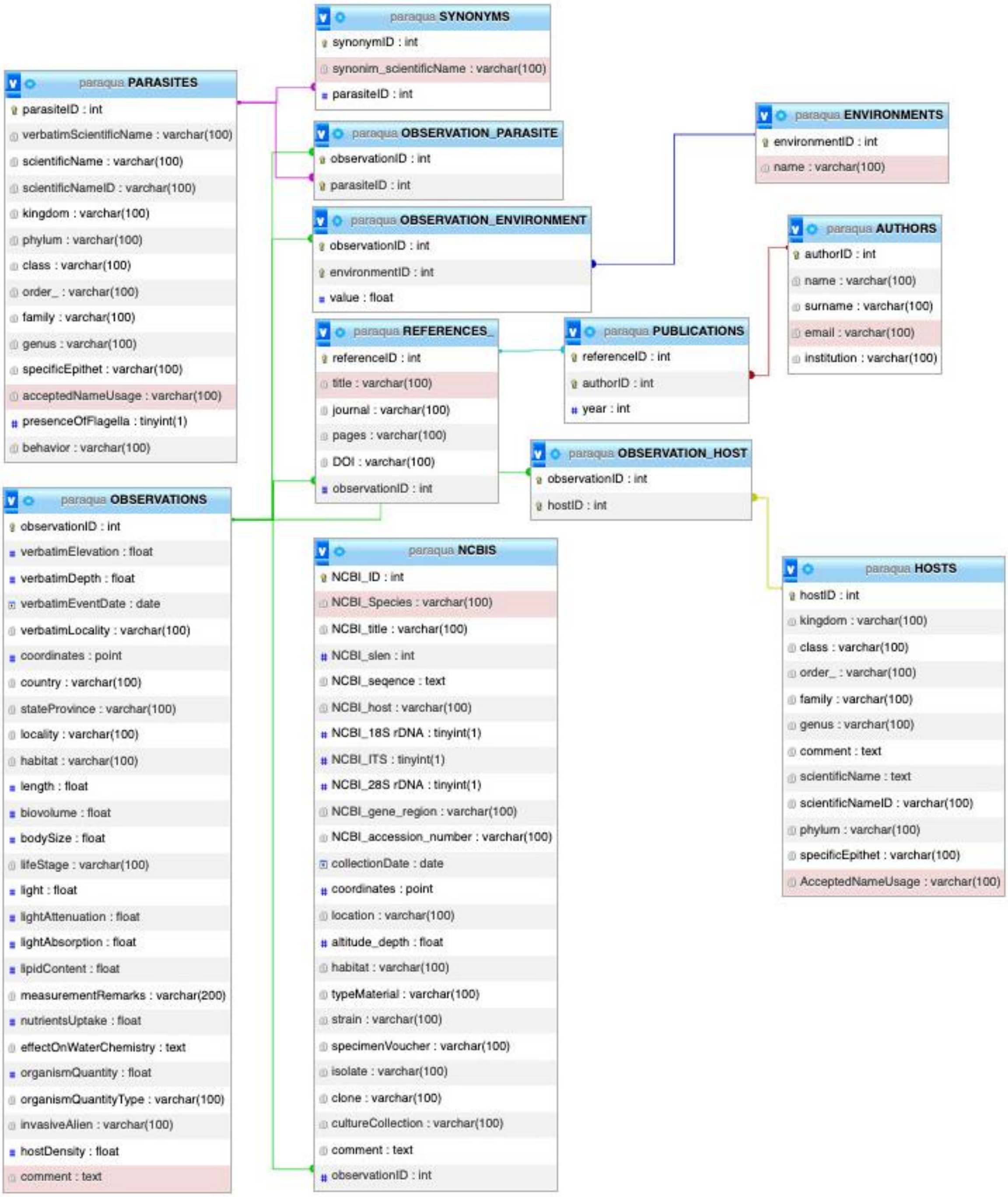

Fig. 4 Database relational schema.

3. Discussion

The collaborative efforts of WG1 and WG2 within the ParAqua Action are, to the best of our knowledge, the first attempt to centralise all data in the field of zoosporic parasites and their interactions with algae. Starting with a rigorous requirements gathering and analysis, a designed survey provided the initial framework for data collection. The response rate was quite satisfactory (35%, 47/135), as it was previously noted (Wu et al., 2022). The subsequent hybrid workshop conducted in Cyprus served as an important occasion for knowledge exchange, enabling the scientific community to align on data management principles. It was a useful way to understand the needs of the scientific community as well.

The adoption of machine-actionable Data Management Plan through the ARGOS platform emphasised our commitment to robust data management practices. This step not only ensured compliance with contemporary standards but also underscored our dedication to transparency and accessibility, cornerstones of credible scientific research. However, we realised that involving the Action partners in this process is not necessarily a good choice. If on one side it empowers researchers on the data management activities, on the other they might not know all the technicalities (such as metadata schemas, semantic resources, etc.), slowing down or somewhat limiting the activity.

During the conceptual design phase, relationships among various entities were delineated through the entity-relationship diagram. This diagram, a visual representation of the complex interconnections within ecological systems, highlighted the nuanced nature of zoosporic parasite interactions. This was possible as we followed the process from the very beginning, and we were able to integrate data coming from diverse sources.

The most difficult step is probably represented by the data mobilisation, i.e. adjusting the datasets obtained from the scientific community in a suitable format to be understood and used in the DB. This activity is still running but is a vital one. Indeed, we can meticulously craft the most exquisite DB with the utmost care, but its true essence only emerges when it's infused with the vital essence of data. Several issues may hamper this activity. First of all, researchers are willing to adapt their datasets to the templates, but they need a considerable support to map the variable with semantic resources, such as controlled vocabularies and thesauri. To overcome this problem, we run this activity by ourselves, without involving the researchers. However, since this activity is hard and time-consuming, it needs dedicated personnel. The mobilisation of data and knowledge is a general challenge within the realms of ecology and biodiversity, especially in terms of FAIRification of data (Dunning et al., 2019). Resolving this issue is fundamental to effectively reuse the amount of data produced by the research.

In the immediate future, our focus will be on the process of completing the collection of data, its harmonisation and, ultimately, its integration into the database. Simultaneously, we will create an intuitive User Interface tailored to facilitate a user-friendly interaction with the DB. This interface will serve as the gateway, empowering researchers, scientists, and lake managers to query the data that will be collected within the ParAqua DB.

Acknowledgement

This publication is based upon work from COST Action 20125 - Applications for zoosporic parasites in aquatic systems (ParAqua), supported by COST (European Cooperation in Science and Technology). GT and DR have been supported by a ParAqua Virtual Mobility grant. The authors are grateful to Serena Rasconi, who provides feedback on the paper.

References

Baker, K. S., Duerr, R. E., & Parsons, M. A. (2015). Scientific knowledge mobilization: Co-evolution of data products and designated communities. International Journal of Digital Curation, 10(2), 110-135.

Carney, L. T., & Lane, T. W. (2014). Parasites in algae mass culture. Front Microbiol 5: 278.

Darwin Core Maintenance Group. 2023. Darwin Core List of Terms. Biodiversity Information Standards (TDWG). http://rs.tdwg.org/dwc/doc/list/2023-09-18

Diack, G., Bull, C., Akenhead, S. A., Van Der Stap, T., Johnson, B. T., Rivot, E., ... & Crozier, W. (2022). Enhancing data mobilisation through a centralised data repository for Atlantic salmon (Salmo salar L.): Providing the resources to promote an ecosystem-based management framework. Ecological Informatics, 70, 101746.

Dunning, A., Sansone, S. A., & Teperek, M. (2019). The layered cake of FAIR Coordination: How many is too many. Scientific Data.

Grami, B., Rasconi, S., Niquil, N., Jobard, M., Saint-Béat, B., & Sime-Ngando, T. (2011). Functional effects of parasites on food web properties during the spring diatom bloom in Lake Pavin: a linear inverse modeling analysis. PloS one, 6(8), e23273.

Naselli-Flores, L., & Padisák, J. (2023). Ecosystem services provided by marine and freshwater phytoplankton. Hydrobiologia, 850(12), 2691-2706.

Rasconi, S., Grossart, H. P., Gsell, A., Ibelings, B. W., Van de Waal, D., Agha, R., ... & Znachor, P. (2022). Applications for zoosporic parasites in aquatic systems (ParAqua). ARPHA Preprints, 3, e94590.

Rasconi, S. (2023). Feedback report from the first ParAqua hybrid meeting-with considerations on challenges and advantages of mixed events. ARPHA Preprints, 4, e102047.

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., ... & Sherry, S. T. (2022). Database resources of the national center for biotechnology information. Nucleic acids research, 50(D1), D20-D26.

Schulman, L., Lahti, K., Piirainen, E., Heikkinen, M., Raitio, O., & Juslén, A. (2021). The Finnish Biodiversity Information Facility as a best-practice model for biodiversity data infrastructures. Scientific data, 8(1), 137.

Tarallo, A., Rosati, I., Garzoli, L., Reñé, A., & Rasconi, S. (2023). ParAqua COST Action CA20125 Data Management Plan (Version 1). Zenodo. https://doi.org/10.5281/zenodo.8154465

Wu, M. J., Zhao, K., & Fils-Aime, F. (2022). Response rates of online surveys in published research: A meta-analysis. Computers in Human Behavior Reports, 7, 100206.

APPENDIX 1. DETAIL OF VARIABLES WITH MAPPING V1.0

Appendix 1. Detail of variables with mapping V1.0

OBSERVATIONS table	
Variable	**Mapping V1**
observationID	N/A
verbatimElevation	Darwin Core
verbatimDepth	Darwin Core
verbatimEventDate	Darwin Core
verbatimLocality	Darwin Core
coordinates	N/D
country	Darwin Core
stateProvince	Darwin Core
locality	Darwin Core
habitat	Darwin Core
length	LifeWatch Trait Thesaurus
biovolume	LifeWatch Trait Thesaurus
bodySize	LifeWatch Trait Thesaurus
lifeStage	LifeWatch Trait Thesaurus
light	N/D
lightAttenuation	N/D
lightAbsorption	N/D
lipidContent	N/D
measurementRemarks	Darwin Core

nutrientsUptake	N/D
effectOnWaterChemistry	N/D
organismQuantity	Darwin Core
organismQuantityType	Darwin Core
invasiveAlien	N/D
hostDensity	N/D
comment	N/D

PARASITES table	
Variable	**Mapping V1**
parasiteID	N/A
verbatimScientificName	Darwin Core
scientificName	Darwin Core
scientificNameID	Darwin Core
kingdom	Darwin Core
phylum	Darwin Core
class	Darwin Core
order	Darwin Core
family	Darwin Core
genus	Darwin Core
specificEpithet	Darwin Core
acceptedNameUsage	Darwin Core
presenceOfFlagella	N/D
behavior	Darwin Core

SYNONYMS table	
Variable	**Mapping V1**
synonymID	N/A
synonim_scientificName	N/D
parasiteID	N/A

HOSTS table	
Variable	**Mapping V1**
hostID	N/A
kingdom	Darwin Core
class	Darwin Core
order_	Darwin Core
family	Darwin Core
genus	Darwin Core
comment	Darwin Core
scientificName	Darwin Core
scientificNameID	Darwin Core
phylum	Darwin Core
specificEpithet	Darwin Core
AcceptedNameUsage	Darwin Core

ENVIRONMENTS table	
Variable	**Mapping V1**
environmentID	N/A
name	N/A

REFERENCES table	
Variable	**Mapping V1**
referenceID	N/A
title	Dublin Core
journal	Dublin Core
pages	Dublin Core
DOI	Dublin Core

AUTHORS table	
Variable	**Mapping V1**
authorID	N/A
name	Dublin Core
surname	Dublin Core
email	Dublin Core
institution	Dublin Core

NCBI table	
Variable	**Mapping V1**
NCBI_ID	N/A
NCBI_Species	NCBI
NCBI_title	NCBI
NCBI_slen	NCBI
NCBI_seqence	NCBI
NCBI_host	NCBI
NCBI_18S rDNA	NCBI
NCBI_ITS	NCBI
NCBI_28S rDNA	NCBI
NCBI_gene_region	NCBI
NCBI_accession_number	NCBI
collectionDate	NCBI
coordinates	NCBI
location	Darwin Core
altitude_depth	NCBI
habitat	Darwin Core
typeMaterial	NCBI
strain	NCBI
specimenVoucher	NCBI
isolate	NCBI
clone	NCBI

cultureCollection	NCBI
comment	N/D
observationID	N/A

Legend:
N/A = Not Available, i.e. for that variable a mapping is not needed as it represents the unique ID which identify uniquely each row in the table
N/D = The mapping is in progress and not yet defined

Appendix 2. DDLs needed for the creation of the tables into the DBMS

```
  --
  -- Database: `paraqua`
  --
  CREATE DATABASE IF NOT EXISTS `paraqua` DEFAULT CHARACTER SET utf8mb4
COLLATE utf8mb4_general_ci;
  USE `paraqua`;

  -- --------------------------------------------------------

  CREATE TABLE `AUTHORS` (
    `authorID` int NOT NULL,
    `name` varchar(100) CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci
DEFAULT NULL,
    `surname` varchar(100) COLLATE utf8mb4_general_ci NOT NULL,
    `email` varchar(100) CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci
DEFAULT NULL,
    `institution` varchar(100) CHARACTER SET utf8mb4 COLLATE
utf8mb4_general_ci DEFAULT NULL
  ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_general_ci;

  -- --------------------------------------------------------

  CREATE TABLE `ENVIRONMENTS` (
    `environmentID` int NOT NULL,
    `name` varchar(100) COLLATE utf8mb4_general_ci NOT NULL
  ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_general_ci;

  -- --------------------------------------------------------

  CREATE TABLE `HOSTS` (
    `hostID` int NOT NULL,
    `kingdom` varchar(100) COLLATE utf8mb4_general_ci DEFAULT NULL,
    `class` varchar(100) COLLATE utf8mb4_general_ci DEFAULT NULL,
    `order_` varchar(100) COLLATE utf8mb4_general_ci DEFAULT NULL,
    `family` varchar(100) COLLATE utf8mb4_general_ci DEFAULT NULL,
    `genus` varchar(100) COLLATE utf8mb4_general_ci DEFAULT NULL,
    `comment` text COLLATE utf8mb4_general_ci,
    `scientificName` text COLLATE utf8mb4_general_ci,
    `scientificNameID` varchar(100) COLLATE utf8mb4_general_ci DEFAULT NULL,
    `phylum` varchar(100) COLLATE utf8mb4_general_ci DEFAULT NULL,
    `specificEpithet` varchar(100) COLLATE utf8mb4_general_ci DEFAULT NULL,
    `AcceptedNameUsage` varchar(100) COLLATE utf8mb4_general_ci DEFAULT
NULL
  ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_general_ci;

  -- --------------------------------------------------------
  CREATE TABLE `NCBIS` (
    `NCBI_ID` int NOT NULL,
    `NCBI_Species` varchar(100) CHARACTER SET utf8mb4 COLLATE
utf8mb4_general_ci DEFAULT NULL,
    `NCBI_title` varchar(1000) CHARACTER SET utf8mb4 COLLATE
utf8mb4_general_ci DEFAULT NULL,
    `NCBI_slen` int DEFAULT NULL,
    `NCBI_seqence` text CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci,
    `NCBI_host` varchar(100) CHARACTER SET utf8mb4 COLLATE
utf8mb4_general_ci DEFAULT NULL,
    `NCBI_18S rDNA` tinyint(1) DEFAULT NULL,
    `NCBI_ITS` tinyint(1) DEFAULT NULL,
    `NCBI_28S rDNA` tinyint(1) DEFAULT NULL,
    `NCBI_gene_region` varchar(100) CHARACTER SET utf8mb4 COLLATE
utf8mb4_general_ci DEFAULT NULL,
    `NCBI_accession_number` varchar(100) CHARACTER SET utf8mb4 COLLATE
utf8mb4_general_ci DEFAULT NULL,
    `collectionDate` date DEFAULT NULL,
    `coordinates` point DEFAULT NULL,
    `location` varchar(100) CHARACTER SET utf8mb4 COLLATE
utf8mb4_general_ci DEFAULT NULL,
    `altitude_depth` float DEFAULT NULL,
    `habitat` varchar(100) CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci
DEFAULT NULL,
    `typeMaterial` varchar(100) CHARACTER SET utf8mb4 COLLATE
utf8mb4_general_ci DEFAULT NULL,
    `strain` varchar(100) CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci
DEFAULT NULL,
    `specimenVoucher` varchar(100) CHARACTER SET utf8mb4 COLLATE
utf8mb4_general_ci DEFAULT NULL,
    `isolate` varchar(100) CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci
DEFAULT NULL,
    `clone` varchar(100) CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci
DEFAULT NULL,
    `cultureCollection` varchar(100) CHARACTER SET utf8mb4 COLLATE
utf8mb4_general_ci DEFAULT NULL,
    `comment` text COLLATE utf8mb4_general_ci,
    `observationID` int NOT NULL
  ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_general_ci;

  -- --------------------------------------------------------

  CREATE TABLE `OBSERVATIONS` (
    `observationID` int NOT NULL,
    `verbatimElevation` float DEFAULT NULL,
    `verbatimDepth` float DEFAULT NULL,
    `verbatimEventDate` date DEFAULT NULL,
    `verbatimLocality` varchar(100) COLLATE utf8mb4_general_ci DEFAULT NULL,
    `coordinates` point DEFAULT NULL,
    `country` varchar(100) COLLATE utf8mb4_general_ci DEFAULT NULL,
    `stateProvince` varchar(100) COLLATE utf8mb4_general_ci DEFAULT NULL,
    `locality` varchar(100) CHARACTER SET utf8mb4 COLLATE
utf8mb4_general_ci DEFAULT NULL,
    `habitat` varchar(100) CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci
DEFAULT NULL,
    `length` float DEFAULT NULL,
    `biovolume` float DEFAULT NULL,
    `bodySize` float DEFAULT NULL,
    `lifeStage` varchar(100) COLLATE utf8mb4_general_ci DEFAULT NULL,
    `light` float DEFAULT NULL,
    `lightAttenuation` float DEFAULT NULL,
    `lightAbsorption` float DEFAULT NULL,
    `lipidContent` float DEFAULT NULL,
    `measurementRemarks` varchar(200) COLLATE utf8mb4_general_ci DEFAULT
NULL,
    `nutrientsUptake` float DEFAULT NULL,
    `effectOnWaterChemistry` text COLLATE utf8mb4_general_ci,
    `organismQuantity` float DEFAULT NULL,
    `organismQuantityType` varchar(100) COLLATE utf8mb4_general_ci DEFAULT
NULL,
    `invasiveAlien` varchar(100) COLLATE utf8mb4_general_ci DEFAULT NULL,
    `hostDensity` float DEFAULT NULL,
    `comment` text CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci
  ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_general_ci;

  -- --------------------------------------------------------

  CREATE TABLE `OBSERVATION_ENVIRONMENT` (
    `observationID` int NOT NULL,
    `environmentID` int NOT NULL,
    `value` float NOT NULL
  ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_general_ci;

  -- --------------------------------------------------------

  CREATE TABLE `OBSERVATION_HOST` (
    `observationID` int NOT NULL,
    `hostID` int NOT NULL
  ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_general_ci;

  -- --------------------------------------------------------

  CREATE TABLE `OBSERVATION_PARASITE` (
    `observationID` int NOT NULL,
    `parasiteID` int NOT NULL
  ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_general_ci;

  -- --------------------------------------------------------

  CREATE TABLE `PARASITES` (
    `parasiteID` int NOT NULL,
    `verbatimScientificName` varchar(100) CHARACTER SET utf8mb4 COLLATE
utf8mb4_general_ci DEFAULT NULL,
    `scientificName` varchar(100) CHARACTER SET utf8mb4 COLLATE
utf8mb4_general_ci DEFAULT NULL,
    `scientificNameID` varchar(100) CHARACTER SET utf8mb4 COLLATE
utf8mb4_general_ci DEFAULT NULL,
    `kingdom` varchar(100) CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci
DEFAULT NULL,
    `phylum` varchar(100) CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci
DEFAULT NULL,
    `class` varchar(100) COLLATE utf8mb4_general_ci DEFAULT NULL,
    `order_` varchar(100) CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci
DEFAULT NULL,
    `family` varchar(100) CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci
DEFAULT NULL,
    `genus` varchar(100) COLLATE utf8mb4_general_ci DEFAULT NULL,
    `specificEpithet` varchar(100) CHARACTER SET utf8mb4 COLLATE
utf8mb4_general_ci DEFAULT NULL,
    `acceptedNameUsage` varchar(100) CHARACTER SET utf8mb4 COLLATE
utf8mb4_general_ci DEFAULT NULL,
    `presenceOfFlagella` tinyint(1) DEFAULT NULL,
    `behavior` varchar(100) COLLATE utf8mb4_general_ci DEFAULT NULL
  ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_general_ci;

  -- --------------------------------------------------------

  CREATE TABLE `PUBLICATIONS` (
    `referenceID` int NOT NULL,
    `authorID` int NOT NULL,
    `year` int NOT NULL
  ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_general_ci;

  -- --------------------------------------------------------

  CREATE TABLE `REFERENCES_` (
    `referenceID` int NOT NULL,
    `title` varchar(1000) CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci
NOT NULL,
    `journal` varchar(100) CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci
DEFAULT NULL,
    `pages` varchar(100) CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci
DEFAULT NULL,
    `DOI` varchar(100) CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci
DEFAULT NULL,
    `observationID` int NOT NULL
  ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_general_ci;

  -- --------------------------------------------------------

  CREATE TABLE `SYNONYMS` (
    `synonymID` int NOT NULL,
    `synonym_scientificName` varchar(100) CHARACTER SET utf8mb4 COLLATE
utf8mb4_general_ci NOT NULL,
    `parasiteID` int NOT NULL
  ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_general_ci;


  ALTER TABLE `AUTHORS`
    ADD PRIMARY KEY (`authorID`);

  ALTER TABLE `ENVIRONMENTS`
    ADD PRIMARY KEY (`environmentID`);

  ALTER TABLE `HOSTS`
    ADD PRIMARY KEY (`hostID`);

  ALTER TABLE `NCBIS`
    ADD PRIMARY KEY (`NCBI_ID`),
    ADD KEY `observationID` (`observationID`);

  ALTER TABLE `OBSERVATIONS`
    ADD PRIMARY KEY (`observationID`);

  ALTER TABLE `OBSERVATION_ENVIRONMENT`
    ADD PRIMARY KEY (`environmentID`,`observationID`),
    ADD KEY `observationID` (`observationID`);

  ALTER TABLE `OBSERVATION_HOST`
    ADD PRIMARY KEY (`hostID`,`observationID`),
    ADD KEY `observationID` (`observationID`);

  ALTER TABLE `OBSERVATION_PARASITE`
```

```
    ADD PRIMARY KEY (`observationID`,`parasiteID`) USING BTREE,
    ADD KEY `occurrenceID` (`parasiteID`);

  ALTER TABLE `PARASITES`
    ADD PRIMARY KEY (`parasiteID`);

  ALTER TABLE `PUBLICATIONS`
    ADD PRIMARY KEY (`referenceID`,`authorID`) USING BTREE,
    ADD KEY `authorID` (`authorID`);

  ALTER TABLE `REFERENCES_`
    ADD PRIMARY KEY (`referenceID`),
    ADD KEY `observationID` (`observationID`);

  ALTER TABLE `SYNONYMS`
    ADD PRIMARY KEY (`synonymID`),
    ADD KEY `occurrenceID` (`parasiteID`);


  ALTER TABLE `AUTHORS`
    MODIFY `authorID` int NOT NULL AUTO_INCREMENT;

  ALTER TABLE `ENVIRONMENTS`
    MODIFY `environmentID` int NOT NULL AUTO_INCREMENT, AUTO_INCREMENT=3;

  ALTER TABLE `HOSTS`
    MODIFY `hostID` int NOT NULL AUTO_INCREMENT;

  ALTER TABLE `NCBIS`
    MODIFY `NCBI_ID` int NOT NULL AUTO_INCREMENT;

  ALTER TABLE `OBSERVATIONS`
    MODIFY `observationID` int NOT NULL AUTO_INCREMENT, AUTO_INCREMENT=2;

  ALTER TABLE `PARASITES`
    MODIFY `parasiteID` int NOT NULL AUTO_INCREMENT;

  ALTER TABLE `REFERENCES_`
    MODIFY `referenceID` int NOT NULL AUTO_INCREMENT;

  ALTER TABLE `SYNONYMS`
    MODIFY `synonymID` int NOT NULL AUTO_INCREMENT;


  ALTER TABLE `NCBIS`
    ADD CONSTRAINT `ncbis_ibfk_1` FOREIGN KEY (`observationID`) REFERENCES
`OBSERVATIONS` (`observationID`) ON DELETE RESTRICT ON UPDATE RESTRICT;

  ALTER TABLE `OBSERVATION_ENVIRONMENT`
    ADD    CONSTRAINT    `observation_environment_ibfk_1`    FOREIGN    KEY
(`observationID`)  REFERENCES  `OBSERVATIONS`  (`observationID`)  ON  DELETE
RESTRICT ON UPDATE RESTRICT,
    ADD    CONSTRAINT    `observation_environment_ibfk_2`    FOREIGN    KEY
(`environmentID`)  REFERENCES  `ENVIRONMENTS`  (`environmentID`)  ON  DELETE
RESTRICT ON UPDATE RESTRICT;

  ALTER TABLE `OBSERVATION_HOST`
    ADD   CONSTRAINT   `observation_host_ibfk_1`   FOREIGN   KEY   (`hostID`)
REFERENCES `HOSTS` (`hostID`) ON DELETE RESTRICT ON UPDATE RESTRICT,
    ADD CONSTRAINT `observation_host_ibfk_2` FOREIGN KEY (`observationID`)
REFERENCES `OBSERVATIONS` (`observationID`) ON DELETE RESTRICT ON UPDATE
RESTRICT;

  ALTER TABLE `OBSERVATION_PARASITE`
    ADD CONSTRAINT `observation_parasite_ibfk_1` FOREIGN KEY
(`observationID`) REFERENCES `OBSERVATIONS` (`observationID`) ON DELETE
RESTRICT ON UPDATE RESTRICT,
    ADD CONSTRAINT `observation_parasite_ibfk_2` FOREIGN KEY (`parasiteID`)
REFERENCES `PARASITES` (`parasiteID`) ON DELETE RESTRICT ON UPDATE RESTRICT;

  ALTER TABLE `PUBLICATIONS`
    ADD CONSTRAINT `publications_ibfk_1` FOREIGN KEY (`referenceID`)
REFERENCES `REFERENCES_` (`referenceID`) ON DELETE RESTRICT ON UPDATE
RESTRICT,
    ADD CONSTRAINT `publications_ibfk_2` FOREIGN KEY (`authorID`)
REFERENCES `AUTHORS` (`authorID`) ON DELETE RESTRICT ON UPDATE RESTRICT;

  ALTER TABLE `REFERENCES_`
    ADD CONSTRAINT `references__ibfk_1` FOREIGN KEY (`observationID`)
REFERENCES `OBSERVATIONS` (`observationID`) ON DELETE RESTRICT ON UPDATE
RESTRICT;

  ALTER TABLE `SYNONYMS`
    ADD CONSTRAINT `synonyms_ibfk_1` FOREIGN KEY (`parasiteID`) REFERENCES
`PARASITES` (`parasiteID`) ON DELETE RESTRICT ON UPDATE RESTRICT;
  COMMIT;
```