Metadata Made Easy - Develop and Use Domain Specific Metadata Schemes by following the dmdScheme approach.

Authors

- **Rainer M Krug**, University of Zürich, Department of Evolutionary Biology and Environmental Studies, Winterthurerstrasse 190, 8057 Zurich, email: Rainer.Krug@uzh.ch, Rainer@Krugs.de
- **Owen L. Petchey**, University of Zürich, Department of Evolutionary Biology and Environmental Studies, Winterthurerstrasse 190, 8057 Zurich, email: Owen.Petchey@ieu.uzh.ch

Corresponding Author

- **Rainer M Krug**, University of Zürich, Department of Evolutionary Biology and Environmental Studies, Winterthurerstrasse 190, 8057 Zurich, email: Rainer.Krug@uzh.ch, Rainer@Krugs.de

Metadata Made Easy - Develop and Use Domain Specific Metadata Schemes by following the dmdScheme Approach.

Rainer M Krug[a], Owen L. Petchey[a]

[a]*Department of Evolutionary Biology and Environmental Studies, Winterthurerstrasse 190, 8057 Zurich*

Abstract

1. Metadata plays an essential role in the long term preservation, reuse, and interoperability of data. Nevertheless, creating useful metadata can be sufficiently difficult and weakly-enough incentivised that many datasets may be accompanied by little or no metadata. One key challenge is, therefore, how to make metadata creation easier and more valuable. We present a solution that involves creating domain specific metadata schemes that are as complex as necessary and as simple as possible. These goals are achieved by co-development between a metadata expert and the researchers (i.e. the data creators). The final product is a bespoke metadata scheme into which researchers can enter information (and validate it) via the simplest of interfaces: a web browser application and a spreadsheet.
2. We provide the R package `dmdScheme` (Krug & Petchey, 2019a) for creating a template domain specific scheme. We describe how to create a domain specific scheme from this template, including the iterative co-development process, and the simple methods for using the scheme, and simple methods for quality assessment, improvement, and validation.
3. The process of developing a metadata scheme following the outlined approach was successful, resulting in a metadata scheme which is used for the data generated in our research group. The validation quickly identifies

Preprint submitted to Ecology and Evolution *February 10, 2021*

Email addresses: Rainer.Krug@uzh.ch \& Rainer@krugs.de (Rainer M Krug), Owen.Petchey@ieu.uzh.ch (Owen L. Petchey)

forgotten metadata, as well as inconsistent metadata, therefore improving the quality of the metadata. Multiple output formats are available, including XML.

4. Making the provision of metadata easier while also ensuring high quality must be a priority for data curation initiatives. We show how both objectives are achieved by close collaboration between metadata experts and researchers to create domain specific schemes. A near-future priority is to provide methods to interface domain specific schemes with general metadata schemes, such as the Ecological Metadata Language, to increase interoperability.

Keywords: metadata quality; data curation; archival; long term storage; R package;

Glossary

- **analysis** - processing the **analysis-ready data** in order to address the research question.
- **analysis-ready data** - data ready for analysis; may be "ready" for a limited set of analyses. An example would be abundance of each of the species in a set of communities (e.g. population dynamic data of ecological communities). (Contrast with **raw data**.)
- **data deposit package** - a collection of data and **metadata** files deposited in a long-term repository. This consists at least one data file and the **rich metadata** describing the data file(s) and associated information. May often contain multiple data files, each with its own **metadata** file.
- **domain / research domain** - a grouping of e.g. experiments, research, and / or questions addressed, whose data sets can be described using **metadata** following one **metadata scheme** which can be regarded as **rich metadata**. One example is "Experimental Microbial Ecology" for which the metadata scheme *emeScheme* (Krug & Petchey, 2019b) was developed. Fields, such as Ecology and Evolutionary Biology, contain numerous domains.
- **domain specific metadata scheme** - a **metadata scheme** for a **domain**.
- **FAIR data principles** - guiding principles to make data Findable, Accessible, Interoperable, and Reusable (Wilkinson et al., 2016).
- **field specific metadata scheme** - a **metadata scheme** general and broad enough to apply to an entire field. E.g. the Ecological Metadata Scheme (EML) (Jones et al., 2019).
- **long-term storage / preservation** - the process of having data stored / preserved and accessible for the long-term (i.e. greater than 20 years

envisaged).

- **long-term (storage) repositories** - repositories which offer **long term storage**. Examples are ('Zenodo - Research. Shared.' 2020) or ('GBIF,' 2020). The Zenodo repository currently has plans defined for at least 20 years of operation.
- **metadata** - data about data. Metadata can be as little as the name of a variable/column in a spreadsheet of data, though such limited metadata would likely not be considered **rich metadata**, and may not make the data FAIR.
- **metadata scheme** - a formalised description of the **metadata** to be included in e.g. a **data deposit package**, their formats, and which ones are compulsory or not. A formal scheme assists with the indexing of the **metadata** that is required for programmatic searching and extracting **metadata** and data from repositories.
- **pre-processing** - the preparation of the **raw data** to make it **analysis-ready**. This should be done by a script to make the process reproducible, and may use different parameters/methods which need to be adjusted based on the research question and the **raw-data**.
- **raw data** - data as provided by the measuring device. This could be images or videos taken from a camera, tables as returned from machines or hand-written records.
- **rich metadata** - defined by the Research Data Alliance (Research Data Alliance, 2017) as "data with enough accurate and relevant attributes to make it easily findable".

Introduction

To define a kind of gold standard for data handling, Wilkinson et al. (2016) developed the so called FAIR data principles. These define principles to make the data **F**indable, **A**ccessible, **I**nteroperable and **R**eusable, and help to assess data handling workflows in regards to openness.

There are multiple reasons why data should be widely reusable (Heaton, 2008; Bishop & Kuula-Luumi, 2017; Pasquetto, Randles, & Borgman, 2017). Widely resusable means that anyone making reasonable efforts could resuse the data, and that this would be the case even if the data creator(s) are unavailable. "Anyone" includes the creator(s) of the data, other members of the creating research group, and any other researcher. Use cases include using data from previous experiments to plan new ones, re-analysing data using different or new pre-processing or analytical approaches to either compare different methodologies (Dufour & Richard, 2019) or to address new scientific aspects (e.g. the use of trait databases Schneider et al. (2019)), meta-analyses (e.g. A. S. Zimmerman, 2008; Culina, Crowther, Ramakers, Gienapp, & Visser, 2018), reproduction of the studies, and use of data for teaching and training (e.g. Atenas & Havemann (2015); or Henty (2015)).

Being able to reuse data includes finding it, understanding why it was collected and how it was generated, understanding which datasets are which, understanding which variables contain what information, and understanding relationships among variables (e.g. Gregory, Groth, Scharnhorst, & Wyatt (2020); Gregory, Groth, Cousijn, Scharnhorst, & Wyatt (2019); A. S. Zimmerman (2008); A. Zimmerman (2007)). All this information should be stored in metadata; thus metadata are essential for reuse (A. Zimmerman, 2007; Gregory, Groth, Scharnhorst, & Wyatt, 2020). Furthermore, interoperability (the I of FAIR) requires standardised metadata schemes.

Metadata schemes have been developed which aim at providing a standardised structure and vocabulary to be used when providing the metadata. Examples of these schemes are Darwin Core (Darwin Core task group, 2014) and the Ecological Metadata Language (short EML) (Jones et al., 2019) in the field of biology / ecology, or more broadly Dublin Core ('Dublin Core,' 2020). Interoperability is essential for research that relies on combining different datasets, and is particularly important for data-based interdisciplinary research as this very often combines data from different sources.

Given such important reasons for accompanying data with appropriate metadata, why do numerous datasets recently published not include useful metadata (Roche, Kruuk, Lanfear, & Binning, 2015)? For example, a search for "ecology" and type "dataset" on the Zenodo website in late June 2020 returned 998 data deposits. The first ten deposits returned contained no deposits with metadata corresponding to any particular metadata scheme. Two deposit contained metadata in a README.md file and a csv file, and one contained a manuscript called metadata. The remainder contained little or no metadata (other than the column names in the datasets, which were not explained in more detail). It may not be far from the truth to say that the majority of thus far deposited datasets, at least on Zenodo in the field of ecology, have so little metadata or such poor quality metadata that they have very little possibility to be reused, at least without considerable effort and with potentially enduring uncertainty.

To have the metadata available requires the producer of the data to provide it. Therefore the answer to the question of why many datasets are deposited without rich metadata is that the data creators have not prioritised creating rich metadata. While there is some interest and some level of prioritisation (e.g. Campbell, Micheli-Campbell, & Udyawer (2018) showed that especially early career researcher are participating in curating and sharing their data and

metadata) a critical question that follows is how to motivate the creation and deposition of appropriate metadata. There are multiple possible answers; one that we focus upon is that creating metadata is not easy, and creating metadata that conforms to a specific scheme is daunting and difficult for researchers. These schemes are relatively complex, as they are not specific to a research domain (see glossary for definition of "research domain"), but rather for a broad field. The advantages of being applicable to a broad field of science (e.g. consistent search across a range a wider range of domains, standardised property names and vocabulary for metadata provision, interoperability) comes with the cost of being complex, somewht complex and rather difficult to understand, which could represent a significant barrier to use by research scientists not working in the field of metadata development.

Our aim was to make the process of creating metadata not only easy, but also useful for the researcher that created the data and, if at all possible, a quite pleasurable experience to create! We follow the suggestion of Poisot, Bruneau, Gonzalez, Gravel, & Peres-Neto (2019), that **domain specific metadata schemes** (small and purpose-built schemes) can be part of the solution to make ecological data easier to find and reuse. The example we use to illustrate a domain specific metadata scheme is from the research domain we term "Experimental Microbial Ecology" (Worsfold, Warren, & Petchey, 2009; e.g. Pennekamp et al., 2017; and Altermatt et al., 2015) (hereafter EME). We illustrate using this domain due to our familiarity with it and because the experimental studies involved can be quite complex. Many measurements are often taken using different methods. Multiple treatments are often applied. Numerous taxa are often involved. Various steps of data processing are required to obtain analysis-ready data (for examples see Pennekamp et al., 2017; and Garnier, Hulot, & Petchey, 2020) from the measured raw data. The methods used can create large amounts

of data (several terabytes). Therefore EME is a sufficiently complex domain to be used as an illustration.

To prevent the proliferation of a multitude of domain specific metadata schemes, risking little or no interoperability among domains, each domain specific scheme should be as much as possible linked formally to standardised metadata schemes. In a sense, a domain specific metadata scheme could be regarded as an easy to use, familiar, intuitive and pleasurable interface to a more general and standardised metadata scheme.

Three other features of domain specific metadata schemes can increase motivation of researchers to use them: co-development, easy of use, and data/metadata validation. Co-development by metadata experts and researchers in respective domains ensures that the scheme can be shaped by providing input to identify essential properties to be included in the metadata, and to exclude non-essential metadata. The goal then is to create a domain specific metadata scheme that fits that domain. Co-development not only results in a better product, but the resulting "ownership" of these schemes by researchers is likely to increase motivation to use them, to advertise them, to provide input for further development, and to include them in teaching and training.

Metadata entry should not be technically difficult, and presumably the easier the better. To accomplish these design goals, we made a metadata entry system that includes only a web browser based application and a spreadsheet. The simplicity of these interfaces should keep the additional workload for the researchers as small as possible. Moreover, these methods of metadata entry can be common across domains, meaning that it is not necessary to teach or learn a different tool for each domain. Previously developed applications for easy metadata entry include Morpho, a data management tool for earth, environmental, and ecological scientists (https://knb.ecoinformatics.org/tools/morpho); it is open source, but

is no longer maintained by the original development team.

Data and metadata validation can help researchers increase the quality of their data and metadata, for example by checking that variables in datasets contain the information they should, and that they correspond to the stated experimental treatment and observations. Most large metadata schemes provide mechanisms for validating the metadata (e.g. EML in the R package EML (Boettiger & Jones, 2019)). These validations assess mainly the syntactical correctness of the metadata, e.g. if all required fields are provided and if numerical values are in the allowed range (if ranges are specified). More detailed (contextual and contentual) validation can be provided for more specific situations or for smaller domains of research, i.e. for domain specific metadata schemes.

In this paper, we present as a case study the experience and results of our research group in developing the EME domain specific metadata scheme. We first used the R package `dmdScheme` (Krug & Petchey, 2019a) to create a template domain specific metadata scheme `emeScheme` (Krug & Petchey, 2019b) and then customised the template scheme to create the EME scheme. We end with a discussion on how these domain specific metadata schemes can be integrated into larger metadata schemes by using the example of EML (Jones et al., 2019).

The content of this article focuses on presenting the approach by which a domain specific metadata scheme can be created using the `dmdScheme` (Krug & Petchey, 2019a) R package, and its advantages in bringing domain specific metadata schemes to more domains and to facilitate the provision of rich and quality assured metadata. This article is supported by two Vignettes: one describes the `dmdScheme` and is aimed at developers of new domain specific schemes and at users interested in a more detailed description of the package. The other vignette is aimed at users of the `emeScheme`, and could be modified for users of other domain specific schemes. Both are included in the Supplementary

information of this article; updated versions are within the respective R packages.

The template dmdScheme Package

The R package `dmdScheme` (Krug & Petchey, 2019a) forms the core of developing and using domain specific metadata schemes following the `dmdScheme` approach. It is normally hidden for the researcher user of the domain specific metadata schemes and mainly of concern for the actual developer of new metadata schemes.

The package contains all the base functionality needed to develop a new domain specific metadata scheme. It includes functionality to create a spreadsheet for entering the domain specific metadata, functionality to read the metadata from that spreadsheet, basic validation functions, and export functions to xml and templates needed to implement the export to EML. It is important to note, that the `dmdScheme` package itself should not be used to enter actual metadata, as it does only contain an example metadata scheme.

How to develop a new scheme and how to use the package is explained in detail in the accompanying vignette Develop and Use the `dmdScheme` which is included in the supplemental material of this article.

A second part of the `dmdScheme` approach is a repository of domain-specific schemes (Krug, 2020). Here any developed domain-specific schemes can be deposited. The R package `dmdScheme` contains functionality to load the selected scheme from this repository and installs the accompanying R package in a temporary library. This arrangement makes it possible to use the scheme not only together with the R package `dmdScheme`, but also in other programming languages, if so desired.

Creating a domain specific metadata scheme

Creating the `emeScheme`

The scheme `emeScheme` (Krug & Petchey, 2019b) was developed based on the `dmdScheme` (Krug & Petchey, 2019a) and is tailored for data from Experimental Microbial Ecology. The motivation to develop this metadata scheme was born out of the realisation that for long-term storage and retrieval following the FAIR data principles, metadata and data format standards are needed to be able to find and retrieve the data at any later stage and to be able to reuse it, even in the own research environment. Therefore it was decided to develop a rich metadata scheme which would provide enough metadata to be able to find the data and to re-use it.

An open exchange between the researchers and a programmer developing the scheme was essential in turning the `emeScheme` into a domain specific metadata scheme which will be used by researchers to create their metadata. Researchers were involved in the process of developing the `emeScheme` from the beginning. This included regular meetings to identify properties in the scheme which are missing, redundant, or not needed. Finally, the researchers were the first testers of the metadata scheme.

The iterative process involved the following steps:

1. Defining the objectives for the scheme. This included the objective of FAIR compliance, but also ease of use and validation functionality.
2. Development of a first version of the scheme. This was done in a spreadsheet which was then imported in the package and forms the basis of the scheme definition.
3. Pilot of entering data from diverse experiments within the domain. The diversity of experiments is important, as different experiments require different metadata properties and even structures.

4. Discussion of experiences of the researchers while entering the metadata, highlighting missing, redundant, or not needed properties in the scheme, etc.
5. Incorporate these into the next revision of the scheme and return to step 2.
6. Finalize the scheme definition package and publish it.

Based on initial discussions, the scheme included information about the experiment itself as well as about the different data sets resulting from different measurements and analysis methods as well as treatments during the experiment. This information about the experiment is valuable contextual metadata. To simplify the provision of the metadata and to avoid duplication of the experimental metadata, all metadata would be entered into one spreadsheet file (with multiple sheets), with any required assignment of metadata to individual datasets done automatically in the final stage of the metadata export.

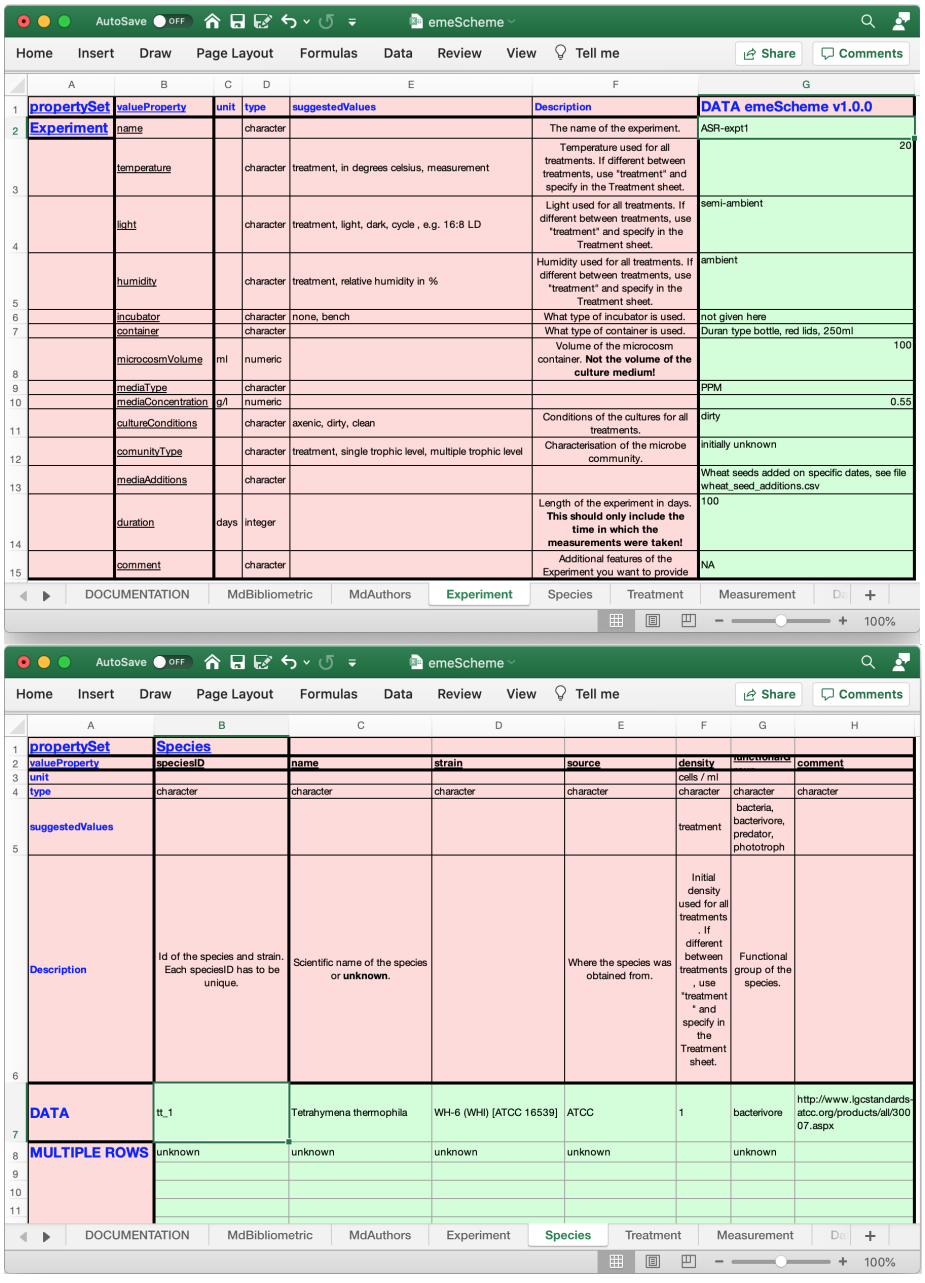
This iterative process resulted in the spreadsheet `emeScheme.xlsx` (Fig. 1 and Supplemental Material).

This scheme was then bundled together with additional examples and uploaded to the dmdScheme repository as emeScheme version 0.9.9 (Krug & Petchey, 2019b).

Enhancing the validation

Even though the package `dmdScheme` already contains a validation function, the validation is generic and mainly structural. The same applies for the export to `xml`, which only exports to a single `xml` file. Additional functionality in the `emeScheme`, i.e. the contextual and contentual validation and the export of the metadata into one `xml` file per data file, is included in an accompanying `emeScheme` R package (Krug & Petchey, 2019b).

Validation means the checking of the internal consistency of the metadata, compliance with the allowed and suggested values and types of the metadata as



propertySet	valueProperty	unit	type	suggestedValues	Description	DATA emeScheme v1.0.0
Experiment	name		character		The name of the experiment.	ASR-expt1
	temperature		character	treatment, in degrees celsius, measurement	Temperature used for all treatments. If different between treatments, use "treatment" and specify in the Treatment sheet.	20
	light		character	treatment, light, dark, cycle , e.g. 16:8 LD	Light used for all treatments. If different between treatments, use "treatment" and specify in the Treatment sheet.	semi-ambient
	humidity		character	treatment, relative humidity in %	Humidity used for all treatments. If different between treatments, use "treatment" and specify in the Treatment sheet.	ambient
	incubator		character	none, bench	What type of incubator is used.	not given here
	container		character		What type of container is used.	Duran type bottle, red lids, 250ml
	microcosmVolume	ml	numeric		Volume of the microcosm container. **Not the volume of the culture medium!**	100
	mediaType		character			PPM
	mediaConcentration	g/l	numeric			0.55
	cultureConditions		character	axenic, dirty, clean	Conditions of the cultures for all treatments.	dirty
	comunityType		character	treatment, single trophic level, multiple trophic level	Characterisation of the microbe community.	initially unknown
	mediaAdditions		character			Wheat seeds added on specific dates, see file wheat_seed_additions.csv
	duration	days	integer		Length of the experiment in days. **This should only include the time in which the measurements were taken!**	100
	comment		character		Additional features of the Experiment you want to provide	NA

propertySet	Species						
valueProperty	speciesID	name	strain	source	density	functionalG	comment
unit					cells / ml		
type	character	character	character	character	character	character	character
suggestedValues					treatment	bacteria, bacterivore, predator, phototroph	
Description	Id of the species and strain. Each speciesID has to be unique.	Scientific name of the species or **unknown**.		Where the species was obtained from.	Initial density used for all treatments . If different between treatments , use "treatment " and specify in the Treatment sheet.	Functional group of the species.	
DATA	tt_1	Tetrahymena thermophila	WH-6 (WH) [ATCC 16539]	ATCC	1	bacterivore	http://www.lgcstandards-atcc.org/products/all/30007.aspx
MULTIPLE ROWS	unknown	unknown	unknown	unknown		unknown	

Figure 1: Two example sheets (Experiment and Species) in the emeScheme metadata file of the ‘emeScheme‘ spreadsheet. The complete spreadsheet can be found in the supplemental material ‘emeScheme.xlsx‘.

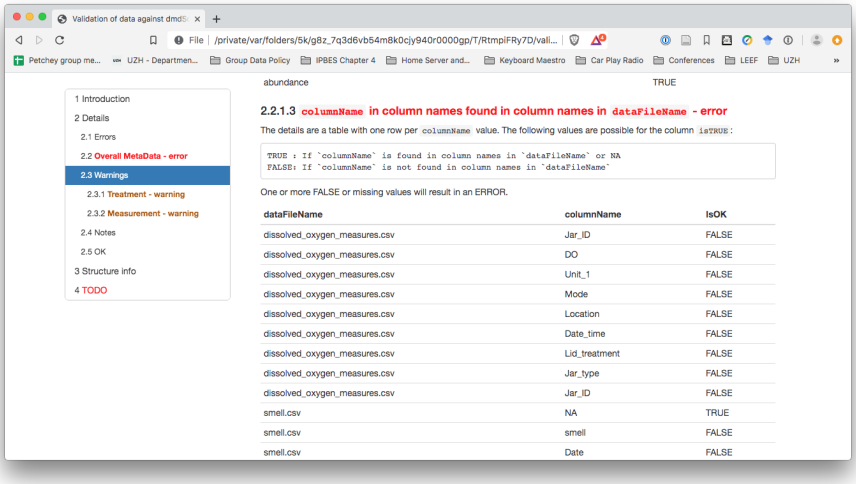


Figure 2: An example of the validation report. The full validation report is in the supplemental material file 'Validation of data against dmdScheme.html'

well as against the structure of the actual data files. This validation produces an `html` (see Fig. 2), `docx` or `pdf` report, which shows errors, warnings or notes. Errors, warnings, and notes represent different levels of severity of detected faults or inconsistencies in the metadata. For example, if a value is not in the list of `allowed values`, it will result in an **error**, while if it is not in the list of `suggested values`, a **note** will be produced. The validation in the `emeScheme` package had to go beyond the validation included in the `dmdScheme` package. Therefore, it was necessary to write a new validation function to add the new validation rules, i.e. the validation of the **structural metadata** which concerns the data files and its columns.

When the validation has completed without errors, the metadata can be exported to one `xml` file per data file. As in the package `dmdScheme` the export to `xml` creates a single `xml` file, and we needed one `xml` file per data file, a new export function was included in the accompanying R package.

Using the `emeScheme`

The functionality in the `emeScheme`, actually of all `dmdScheme` derived metadata schemes, can be accessed by any of three approaches. As the scheme (and the accompanying R package) can be uploaded to the scheme repository (Krug, 2020), they are usable from a universal web app (Krug, 2019) (Fig. 3). Each time the web app is started, it re-loads a list of available scheme packages (and their accompanying R packages), and these can then be used in the app.

Even though this approach is the easiest, it requires the uploading of the metadata as well as the data to the server for validation. This might not be feasible because of confidentiality / privacy reasons or because of the large size of the data files. In this case, the app can also be launched from a local R session. The app then runs on the local computer and data never leaves the local computer.

As a third option, the `emeScheme` and all `dmdScheme` derived packages can also be used from the R command line.

The different approaches of how this can be done are explained in detail in the supplemental material *Develop and Use the dmdScheme*. In addition, the document includes detailed information on how new schemes can be developed. A more hands on user oriented working example of the `emeScheme` is in the supplemental material *emeScheme User Manual*.

Integration of Domain Specific Metadata Schemes into the EML Landscape

As mentioned in the Introduction section, interoperability across domains requires common cross-domain metadata languages. The `dmdScheme` package contains the basic structures to provide an export to EML `xml` format. One of the basic requirements of doing so is linking of the domain specific metadata

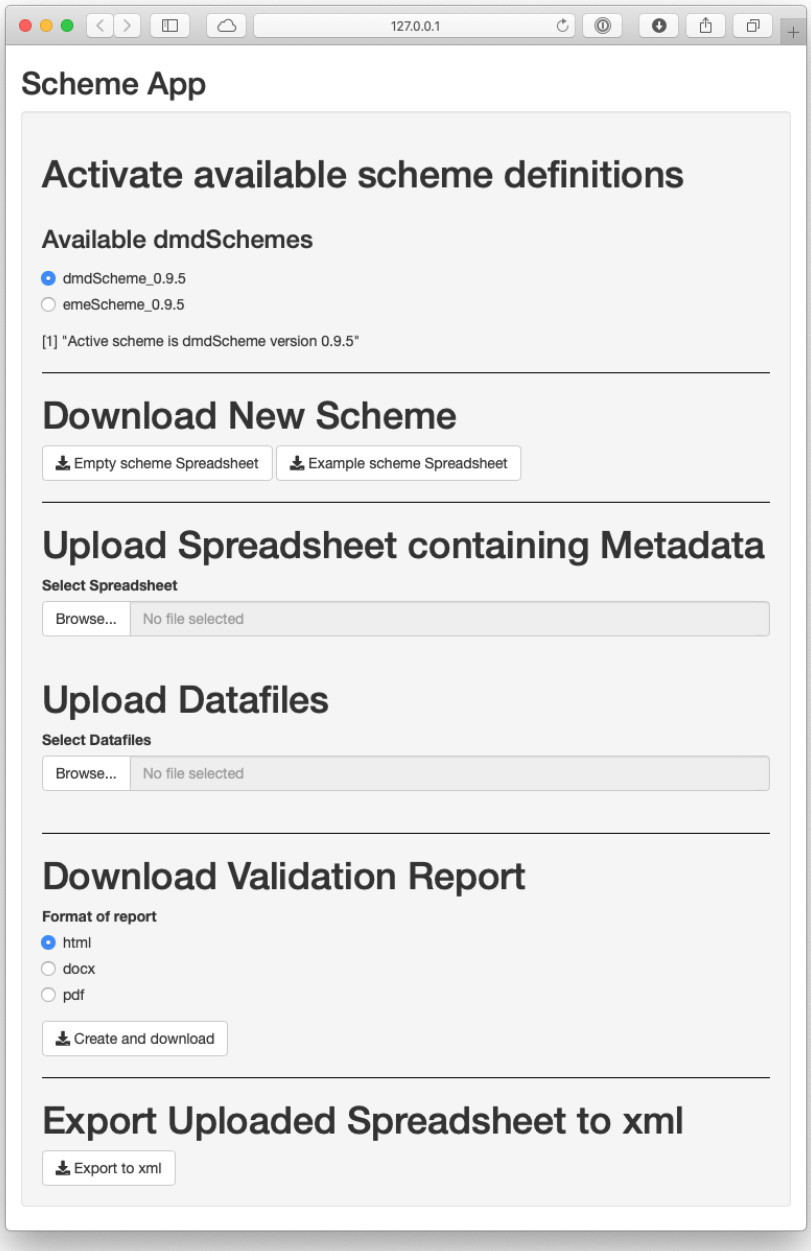


Figure 3: Web app to use the functionality of ‘dmdScheme‘ derived metadata schemes. This app can be run as a universal web app or also locally.

properties to the EML properties. Hence, close inspection of the `emeScheme` (Krug & Petchey, 2019b) and some additional constraints (i.e. only one measurement and extraction method per datafile), make it possible to translate the `emeScheme` metadata into EML. The export into EML is planned for the next major release of the `emeScheme` package.

Export to EML opens a new use case of the `dmdScheme`: if during the development of a new `dmdScheme` the EML scheme is kept in mind, it will be possible to use all the functionalities of the package `dmdScheme` as a frontend for providing EML compliant metadata. In the same way, other large metadata schemes could be used as the framework for the domain specific metadata schemes. This would bridge the gap between simple to understand domain specific metadata schemes on the one side and complex and difficult to understand but applicable to a large range of different domains metadata schemes.

Authors Contributions

RMK and OLP conceived the ideas and designed the methodology; RMK implemented the ideas in R; OLP reviewed the workflow and was the first tester of the package; RMK and OLP led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Data Accessability

The package does not use any data. The code is available as followed:

- dmdScheme Package: The package is available on github at https://github.com/Exp-Micro-Ecol-Hub/dmdScheme . The version used in this paper (v1.2) has the doi https://doi.org/10.5281/zenodo.3894237
- emeScheme Package: The package is available on github at https://github.com/Exp-Micro-Ecol-Hub/emeScheme . The version used in this paper (v1.1.7) has the doi https://doi.org/10.5281/zenodo.4529180

Achknowledgements

We have to thank all members of the Predictive ecology Group at the University of Zurich who provided input in the development of the emeScheme and functioned as guinea pigs in developing and testing this approach.

The authors declare that they have no conflict of interest.

Funding was provided by the SNF Project 310030_188431, and University Research Priority Programme Global Change and Biodiversity.

References

Altermatt, F., Fronhofer, E., Garnier, A., Giometto, A., Hammes, F., Klecka, J., ... Petchey, O. L. (2015). Big answers from small worlds: A user's guide for protist microcosms as a model system in ecology and evolution. *Methods in Ecology and Evolution*, *6*(2), 218–231. doi:10.1111/2041-210X.12312

Atenas, J., & Havemann, L. (2015, November). Open Data as Open Educational Resources: Case Studies of Emerging Practice. doi:10.6084/m9.figshare.1590031.v1

Bishop, L., & Kuula-Luumi, A. (2017). Revisiting Qualitative Data Reuse: A Decade On. *SAGE Open*, *7*(1), 2158244016685136. doi:10.1177/2158244016685136

Boettiger, C., & Jones, M. B. (2019). *EML: Read and write ecological metadata language files*.

Campbell, H. A., Micheli-Campbell, M. A., & Udyawer, V. (2018). Early Career Researchers Embrace Data Sharing. *Trends in Ecology & Evolution*. doi:10.1016/j.tree.2018.11.010

Culina, A., Crowther, T. W., Ramakers, J. J. C., Gienapp, P., & Visser, M. E. (2018). How to do meta-analysis of open datasets. *Nature Ecology & Evolution*, *2*(7), 1053–1056. doi:10.1038/s41559-018-0579-2

Darwin Core task group, B. I. S. (TDWG). (2014, November). Darwin Core: 2014-11-08. Zenodo. doi:10.5281/ZENODO.12694

Dublin Core: Metadata Terms. (2020). https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

Dufour, I. F., & Richard, M.-C. (2019). Theorizing from secondary qualitative data: A comparison of two data analysis methods. *Cogent Education*, *6*(1), 1690265. doi:10.1080/2331186X.2019.1690265

Garnier, A., Hulot, F. D., & Petchey, O. L. (2020). Manipulating the strength of organismenvironment feedback increases nonlinearity and apparent hysteresis of ecosystem response to environmental change. *Ecology and Evolution*,

10(12), 5527–5543. doi:10.1002/ece3.6294

GBIF. (2020). https://www.gbif.org/.

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2019). Searching Data: A Review of Observational Data Retrieval Practices in Selected Disciplines. *Journal of the Association for Information Science and Technology*, *70*(5), 419–432. doi:10.1002/asi.24165

Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Lost or Found? Discovering Data Needed for Research. *Harvard Data Science Review*. doi:10.1162/99608f92.e38165eb

Heaton, J. (2008). Secondary Analysis of Qualitative Data: An Overview. *Historical Social Research / Historische Sozialforschung*, *33*(3 (125)), 33–45.

Henty, M. (2015). *Teaching with Research Data* (Report to the {{Australian National Data Service}} ({{ANDS}})) (p. 37).

Jones, M., O'Brien, M., Mecum, B., Boettiger, C., Schildhauer, M., Maier, M., ... Chong, S. (2019). Ecological metadata language version 2.2.0. doi:10.5063/f11834t2

Krug, R. M. (2019). dmdScheme App. https://rmkrug.shinyapps.io/dmd_app/.

Krug, R. M. (2020, January). dmdSchemeRepository. https://github.com/Exp-Micro-Ecol-Hub/dmdSchemeRepository.

Krug, R. M., & Petchey, O. L. (2019a). dmdScheme: An r package for working with domain specific MetaData schemes (Version v0.9.22). doi:10.5281/zenodo.3581970

Krug, R. M., & Petchey, O. L. (2019b). *emeScheme: A package for working with ecological microbial experimental MetaData*. Zenodo. doi:10.5281/zenodo.1544945

Pasquetto, I., Randles, B., & Borgman, C. (2017). On the Reuse of Scientific Data. *Data Science Journal*, *16*(0), 8. doi:10.5334/dsj-2017-008

Pennekamp, F., Griffiths, J. I., Fronhofer, E. A., Garnier, A., Seymour, M., Altermatt, F., & Petchey, O. L. (2017). Dynamic species classification of

430 microorganisms across time, abiotic and biotic environmentsA sliding window
431 approach. *PLOS ONE*, *12*(5), e0176682. doi:10.1371/journal.pone.0176682
432 Poisot, T., Bruneau, A., Gonzalez, A., Gravel, D., & Peres-Neto, P. (2019).
433 Ecological Data Should Not Be So Hard to Find and Reuse. *Trends in*
434 *Ecology & Evolution*, *0*(0). doi:10.1016/j.tree.2019.04.005
435 Research Data Alliance. (2017). Rich Metadata - DFT. https://smw-rda.esc.rzg.mpg.de/index.php/Rich_Metada
436 Roche, D. G., Kruuk, L. E. B., Lanfear, R., & Binning, S. A. (2015). Public
437 Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLOS*
438 *Biology*, *13*(11), e1002295. doi:10.1371/journal.pbio.1002295
439 Schneider, F. D., Fichtmueller, D., Gossner, M. M., Güntsch, A., Jochum,
440 M., König-Ries, B., ... Simons, N. K. (2019). Towards an ecological
441 trait-data standard. *Methods in Ecology and Evolution*, *10*(12), 2006–2019.
442 doi:10.1111/2041-210X.13288
443 Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton,
444 M., Baak, A., ... Mons, B. (2016). The FAIR Guiding Principles for
445 scientific data management and stewardship. *Scientific Data*, *3*, 160018.
446 doi:10.1038/sdata.2016.18
447 Worsfold, N. T., Warren, P. H., & Petchey, O. L. (2009). Context-dependent
448 effects of predator removal from experimental microcosm communities. *Oikos*,
449 *118*(9), 1319–1326. doi:10.1111/j.1600-0706.2009.17500.x
450 Zenodo - Research. Shared. (2020). https://zenodo.org/.
451 Zimmerman, A. (2007). Not by metadata alone: The use of diverse forms of
452 knowledge to locate data for reuse. *International Journal on Digital Libraries*,
453 *7*(1-2), 5–16. doi:10.1007/s00799-007-0015-8
454 Zimmerman, A. S. (2008). New Knowledge from Old Data: The Role of
455 Standards in the Sharing and Reuse of Ecological Data. *Science, Technology,*
456 *& Human Values*, *33*(5), 631–652. doi:10.1177/0162243907306704