

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

Rainer Krug¹ and Owen Petchey¹

¹University of Zurich Faculty of Mathematics and Science

February 15, 2021

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']` (<https://CRAN.R-project.org/package=dmdScheme>) [Krug2019] 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 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 very 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.

Hosted file

CompleteManuscript.pdf available at <https://authorea.com/users/395951/articles/509066-metadata-made-easy-develop-and-use-domain-specific-metadata-schemes-by-following-the-dmdscheme-approach>

emeScheme							
Home Insert Draw Page Layout Formulas Data Review View Tell me Share Comments							
	A	B	C	D	E	F	G
1	propertySet	valueProperty	unit	type	suggestedValues	Description	DATA emeScheme v1.0.0
2	Experiment	name	character			The name of the experiment.	ASR-exp1
3		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
4		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
5		humidity	character	treatment, relative humidity in %		Humidity used for all treatments. If different between treatments, use "treatment" and specify in the Treatment sheet.	ambient
6		incubator	character	none, bench		What type of incubator is used.	not given here
7		container	character			What type of container is used.	Duran type bottle, red lids, 250ml
8		microcosmVolume	ml	numeric		Volume of the microcosm container. Not the volume of the culture medium!	100
9		mediaType	character				PPM
10		mediaConcentration	g/l	numeric			0.55
11		cultureConditions	character	axenic, dirty, clean		Conditions of the cultures for all treatments.	dirty
12		communityType	character	treatment, single trophic level, multiple trophic level		Characterisation of the microbe community.	initially unknown
13		mediaAdditions	character			Wheat seeds added on specific dates, see file wheat_seed_additions.csv	
14		duration	days	integer		Length of the experiment in days. This should only include the time in which the measurements were taken!	100
15		comment	character			Additional features of the Experiment you want to provide	NA
DOCUMENTATION MdBibliometric MdAuthors Experiment Species Treatment Measurement 100%							
emeScheme							
Home Insert Draw Page Layout Formulas Data Review View Tell me Share Comments							
	A	B	C	D	E	F	G
1	propertySet	Species					
2	valueProperty	speciesID	name	strain	source	density	comment
3	unit	character	character	character	character	cells / ml	character
4	type	character	character	character	character	character	character
5	suggestedValues					treatment	bacteria, bacterivore, predator, phototroph
6	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.
7	DATA	tl_1	Tetrahymena thermophila	WH-6 (WH) [ATCC 16539]	ATCC	1	bacterivore
8	MULTIPLE ROWS	unknown	unknown	unknown	unknown	unknown	http://www.lgcstandards.atcc.org/products/all/30007.aspx
9							
10							
11							
DOCUMENTATION MdBibliometric MdAuthors Experiment Species Treatment Measurement 100%							

1 Introduction

2 Details

2.1 Errors

2.2 Overall MetaData - error

2.3 Warnings

2.3.1 Treatment - warning

2.3.2 Measurement - warning

2.4 Notes

2.5 OK

3 Structure info

4 TODO

abundance

TRUE

2.2.1.3 columnName in column names found in column names in dataFileName - error

The details are a table with one row per columnName value. The following values are possible for the column isTRUE:

isTRUE : If "columnName" is found in column names in "dataFileName" or NA
FALSE: If "columnName" is not found in column names in "dataFileName"

One or more FALSE or missing values will result in an ERROR.

dataFileName	columnName	isOK
dissolved_oxygen_measures.csv	Jar_ID	FALSE
dissolved_oxygen_measures.csv	DO	FALSE
dissolved_oxygen_measures.csv	Unit_1	FALSE
dissolved_oxygen_measures.csv	Mode	FALSE
dissolved_oxygen_measures.csv	Location	FALSE
dissolved_oxygen_measures.csv	Date_time	FALSE
dissolved_oxygen_measures.csv	Lid_treatment	FALSE
dissolved_oxygen_measures.csv	Jar_type	FALSE
dissolved_oxygen_measures.csv	Jar_ID	FALSE
smell.csv	NA	TRUE
smell.csv	smell	FALSE
smell.csv	Date	FALSE

Scheme App

Activate available scheme definitions

Available dmdSchemes

☒ dmdScheme_0.9.5

☐ emeScheme_0.9.5

[1] *Active scheme is dmdScheme version 0.9.5"

Download New Scheme

Empty scheme Spreadsheet

Example scheme Spreadsheet

Upload Spreadsheet containing Metadata

Select Spreadsheet

Browse...

No file selected

Upload Datafiles

Select Datafiles

Browse...

No file selected

Download Validation Report

Format of report

☒ html

☐ docx

☐ pdf

Create and download

Export Uploaded Spreadsheet to xml

Export to xml