Project Title **Implementing FAIR Workflows: a proof of concept study in the field of consciousness**

TWCF Grant No **TWCF0568**

D2.2 Metadata template development for cognitive neuroscience research

Date **28. 02. 2023**

Authors **Xiaoli Chen** https://orcid.org/0000-0003-0207-2705
Kelly Stathis https://orcid.org/0000-0001-6133-4045
Lucia Melloni https://orcid.org/0000-0001-8743-5071
Helena Cousijn https://orcid.org/0000-0001-6660-6214
Zefan Zheng https://orcid.org/0000-0001-9631-7560

Status **Public**

TEMPLETON WORLD
CHARITY FOUNDATION

This project was made possible through the support of a grant from Templeton World Charity Foundation, Inc. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of Templeton World Charity Foundation, Inc.

Executive summary

The Implementing FAIR Workflows Project aims to leverage existing persistent identifier infrastructure, research tools, and platforms to build a proof of concept research workflow for neuroscience research that is FAIR on inception. The project will provide an exemplar workflow to enable and encourage the wider neuroscience community to adopt FAIR practices.

Making research outputs FAIR and traceable can be accomplished by applying the correct persistent identifiers (PIDs) and creating rich and complete metadata. However, metadata schemas for PIDs, such as the DataCite metadata schema, are kept discipline-agnostic to ensure wide applicability across domains and resource types. Domain-specific metadata serves research communities with increased discoverability by providing more granular and more relevant facets of data and requires a different, integrated solution.

As part of the effort to develop domain-specific metadata to make outputs FAIR by adopting semantic standardization and compliance with research community recommendations, the project team has invested in defining and developing a metadata template for cognitive neuroscience studies with human subjects, using the CEDAR Workbench. This document outlines the background, setup, tools used, the process for metadata definition, template building, testing, and community feedback. The template in its current form is presented, along with an overview of the challenges encountered while implementing the template. Future steps entail plans to improve the domain-specific metadata specifications and further evaluate domain ontology building and adoption. We have also identified areas that need improvement, including ease of use of the template interface, and approach to consensus building in the community. Those key future steps will keep the template relevant and useful to the community long term.

Table of Content

Introduction

FAIR Workflows project

The Implementing FAIR workflows project aims to build an exemplar workflow for researchers and research tools by working with a Cognitive Neuroscience research team to implement FAIR practices throughout a 3 year research project. Leveraging existing services and platforms that are integrated with open persistent identifier (PID) infrastructures, the project charts the way forward for researchers and research teams that are looking to take action towards making their research more reproducible and reusable, as well as potential integrators that are interested in providing support to research workflows and make concrete contributions to the scholarly infrastructure on a metadata level.

The project examines every stage of the research process and identifies gaps where FAIR practices can be implemented to improve the FAIRness of the research outputs. The consistent capturing of metadata on different aspects, levels, and granularity is at the core of rendering outputs FAIR (Musen et al., 2022). Creating a domain-specific metadata template directly responds to the interoperability and reusability assessment criteria (Devaraju, Anusuriya et al., 2022):

- Metadata follows a standard recommended by the target research community of the data.
- Metadata is represented using a formal knowledge representation language.
- Metadata uses semantic resources.

Domain-specific Metadata

The neuroscience community has called for standardization of neuroscience data (Poline et al., 2022) and a plethora of efforts have been made across the field, on various levels of data for different use cases. Community initiatives for data standardization like BIDS (Gorgolewski et al., 2016) and the Neuroimaging Data Model (NIDM)[1] provided the groundwork for an agreed-upon vocabulary, a convention for file structuring, and description for datasets.

In the project, we surveyed the efforts around data standardization, data sharing platforms, and ontologies available in the biosciences, particularly the consciousness research domain to enable the development of a domain-specific metadata template that will be used in this project.

CEDAR Template

The CEDAR Workbench (referred to as CEDAR in the rest of the text) is the tool of choice for template building. As a project partner, the CEDAR team has supported the template building

[1] https://www.incf.org/wg/incf-working-group-nidm

process. In addition, they will complete the required development of the CEDAR editable template component to ensure the successful implementation of CEDAR templates on Dryad.

The CEDAR template workbench translates the metadata schema into a fillable form. Through CEDAR, researchers can enter semantically precise controlled terms, and in doing so contribute to a shared understanding of the semantic relations within a specific domain. CEDAR also supports a highly interoperable JSON-based model for templates and metadata, and an API built for flexible integration with external systems.[2]

Dryad Integration

Dryad is a generic purpose research data repository platform that supports data publishing in conjunction with journal article publishing workflows. Researchers can submit a dataset and analysis code used to produce the results reported in an article to Dryad, and associate the dataset with the article through the metadata of the dataset.

The Dryad dataset publishing workflow includes a data submission form that asks the researcher to provide metadata for the dataset, including the title, author(s), relation to manuscript research domain, facility, and funding information.

Dryad will implement the domain specific metadata template once it is complete, to present it as an option for the neuroscience community to supplement the general metadata already collected by Dryad.

> Dryad will integrate the CEDAR Embeddable Editor component to allow researchers to select and fill out appropriate metadata templates during the data deposition process. The discipline-specific metadata created will be saved and made publicly available alongside the datasets being deposited. When a template is selected by a researcher to accompany the data deposition, Dryad will include the DOI of the template in the DOI metadata of the dataset as a related identifier.

Box 1. Excerpt from Dryad project contribution schedule of work.

Summary

For the project, the researchers defined and selected a series of domain-specific metadata to then develop a metadata template for cognitive neuroscience research using the CEDAR workbench, with the goal of promoting the standardization and reusability of metadata in this

[2] https://metadatacenter.github.io/cedar-manual/faq

domain. This template captures information such as study design, experimental setup, data modality, and data sources, and is designed to conform to the FAIR principles of findability, accessibility, interoperability, and reusability.

The template will be implemented on Dryad, a repository for data underlying scientific publications, to enable researchers to deposit their data along with the standardized metadata. This will not only promote the reuse of the data and facilitate discovery and citation of the underlying research, but also promote collaboration in the field of consciousness neuroscience research.

Approach

Preparation of the metadata template in CEDAR

To prepare for the domain-specific metadata work, the research team attended the Metadata For Machines (MFM) workshop[3] to familiarize themselves with best practices for creating FAIR metadata and metadata templates that are machine-actionable.

The project team also surveyed on-going projects in neuroinformatics that are undertaking metadata standard building efforts through the INCF community, and collected input from the project advisory group at the beginning of the work package as reference for work plan and set priorities. One of the key recommendations derived from the desk research and advisory group was to establish a clear use case and not to "reinvent the wheel" and use existing community resources whenever possible.

Exploring use cases of domain specific metadata

Metadata is useful on different levels, in different scenarios, and for different user groups. Metadata schemas used for persistent identifiers generally stay high-level and domain-agnostic to accommodate use cases across the research landscape and provide infrastructural support for connection and interoperability. Domain-specific metadata, on the other hand, will need to take into consideration the current and desired research activities arranged around data creation, use (and reuse), management, and sharing within a specific community.

The initial hypotheses the research team worked on were that 1) the field of consciousness research lacks a domain specific ontology that a metadata schema can draw on, and 2) that traditionally there has not been a workflow to capture the key metadata elements in the process of the research. For these, the research team explored two directions: ontology development and comprehensive metadata descriptor development (which was undertaken together with a parallel project).

[3] https://www.go-fair.org/resources/go-fair-workshop-series/metadata-for-machines-workshops/

For the former, the researchers conducted an exploratory study on keywords used in cognitive neuroscience research publications. To explore the corpus of terminologies used in identifying research publications in the consciousness domain, keywords from two main research journals in the domain were extracted using Web of Science. The resulting dataset was analyzed using the VOSViewer[4] (Figure 1). The research team proposed a work plan to derive a domain ontology based on the keyword mapping; however, this is considered not feasible within the FAIR Workflows project timeline and would require a follow up project.

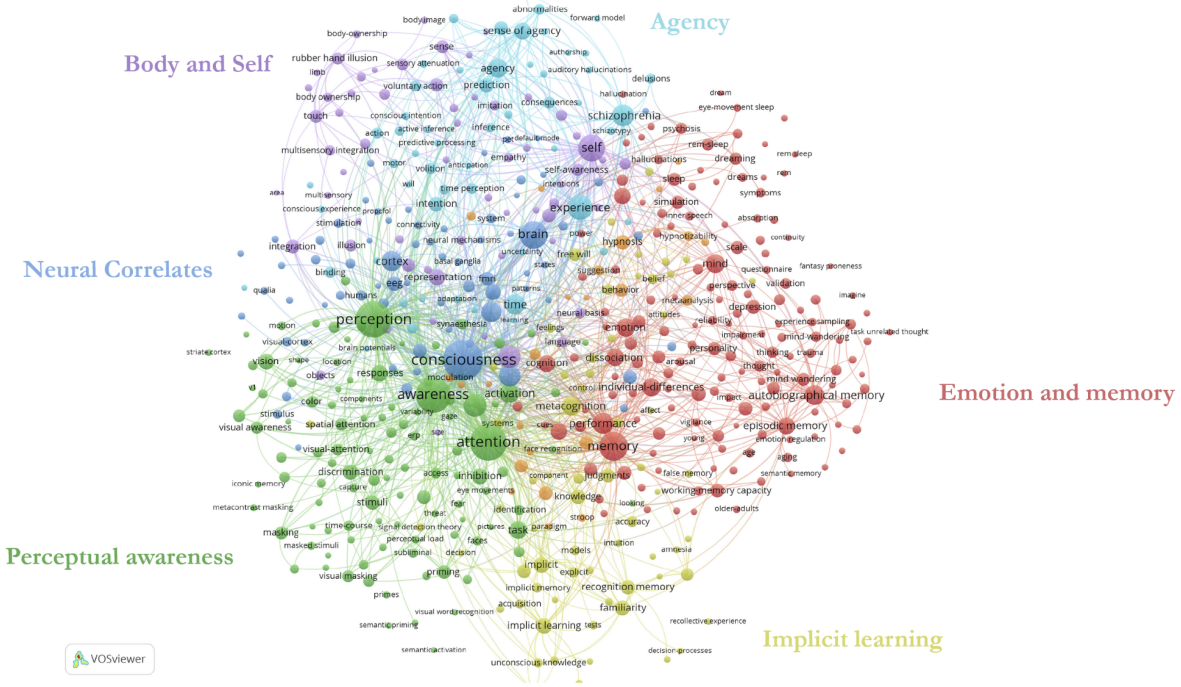


Figure 1. Cluster-based analysis of all collected keywords from neuroscience research publications

Moving forward with the template building, the research team focused on selecting the domaim specific metadata elements to include in the template. The research team included elements that are used to differentiate datasets by data modality and acquisition technique, experimental paradigm, state of preprocessing, etc. without overlapping with detailed descriptive metadata supported by community endorsed data structures such as BIDS. In synergy with a parallel effort for another TWCF funded project (Arc-Cogitate) at the MPIEA, the research team used the GEneric NEuro MEtadata DEscriptor (GENEMEDE) framework as a basis, compared with the data model used on other existing domain-specific data repositories to build a suite of elements that is able to minimally describe the dataset, the experiment setup, and analysis conditions leading to the

[4] https://www.vosviewer.com/

production of the dataset. The selection of the domain-specific metadata fields to be included in the metadata template was largely drawn by identifying a 'minimal' set that will improve data reuse and discoverability while also increasing the chances that researchers will adopt the metadata template usage. An important consideration was the lack of incentives for researchers to provide all relevant metadata. The current implementation then aimed at striking a balance that facilitates adoption, and this imposes minimal requirements. It is thus viewed as incomplete, but as a first step towards incentivizing the research community.

Comparing across platforms

To decide on the core set of properties to include in the template, five data description models (metadata required on different platforms) were compared (Table 1), to 1) include key properties used in domain specific platforms, 2) identify gaps based on research team's requirements, and 3) exclude generic properties that are already covered in the Dryad (color-coded red in the table).

Platform	**Type**	**Domain**
Dryad	Data repository	Generic
GENEMEDE	Data description model	Neuroscience
OpenNeuro	Data repository	Neuroscience
Dandi	Data repository	Neurophysiology
BIDS dataset description	Data description model	Neuroimaging

Table 1. The five models used in this comparative study

A first draft of the template was created with a set of elements (color-coded blue) compiled from the comparative exercise (Table 2). The initial set of elements included contributor role, data modality, subject (number of subject and species), ethics approval, data standard and derivation, experimental design, and related information (study, project, metadata). This list was further divided into five categories: contributor, data, experiment, analysis, and other.

Property	**Dryad**	**GENEMEDE**	**OpenNeuro**	**Dandi**	**BIDS dataset description**	**Properties to include in CEDAR template**
title	dataset title		dataset name	dandiset title	name	/
author	author(s)	researcher	dataset uploader	Contact	Authors	/
				dandiset contributors		contributor role
				Acknowledgements	Acknowledgements	

					How to acknowledge	
domain specific	research domain		domain investigated	study target		/
		device	modality			modality
			species	species		species
			diagnoses status			
			task completed			
			participant age range			
				ethics approval	Ethics approvals	ethics approval
			number of trials			
				number of samples		
				number of subjects		number of subjects
				subject matter of the dataset		
		subject				
		session				
		study				related study
methods	methods		study design	Approach		/
				variables measured		attributes measured
		procedure		measurement technique		availability of protocols
			longitudinal?			longitudinal?
		experiment set up				
funding information	funding				Funding	/
general information of upload	abstract			Description		
			Dataset derivatives availability		Dataset type	Dataset derivatives availability
				Data Standard		Data standard
				number of files		
keyword	keyword			keywords		/
license				License	License	
	usage notes					

related works	related works		associated OpenNeuro dataset	related resources	References and links	/
			data publications			
					Source datasets	
						link to full metadata/ related metadata
related project		Project		associated projects	Generated by	related project
platform specific	relation to paper					
			OpenNeuro Dataset ID		BIDS version	
			Dataset URL		Dataset links	
			Dataset DOI		Dataset DOI	
			published date		HED Version	
			most recent snapshot			
institute/ organization	research facility	lab				

Table 2. Property comparison across platforms

After the elements were selected, ontological resources in the domain of neuroscience and specifically cognitive research were surveyed and included in the draft template as a proof of concept. We particularly prioritized ontologies used by established data models (data models that are actively used, supported, endorsed, and/or sustained by the domain community).

Terms referenced in the template are sourced from these ontologies:

1. Cognitive Atlas Ontology
2. Cognitive Paradigm Ontology
3. Neuroscience Information Framework (NIF) Standard Ontology (NIFSTD)
4. NIDM-Results
5. Statistics Ontology
6. Software Ontology
7. National Cancer Institute Thesaurus

Other ontologies consulted but not used in the template (in its current form) include:

8. Cognitive Reserves Assessment Tasks
9. Computational Neuroscience Ontology
10. Foundational Model of Anatomy
11. Neuroscience Domain Insight Graph

12. Provenance Ontology

The draft template was then used as the basis for an in-person template workshop held in Frankfurt at the end of 2022.

Workshop and testings

The project team arranged a workshop in Frankfurt at the MPI to work with a group of neuroscience researchers. The goals of the workshop were to:

1. Agree on the elements to include in the template
2. Generate descriptions and help text for each elements
3. Testing the usability of template

A draft template was presented to provide a basis for constructive feedback and improvement.

Miro board exercises were used to guide the group through the collaborative process of element selection, and prioritization.

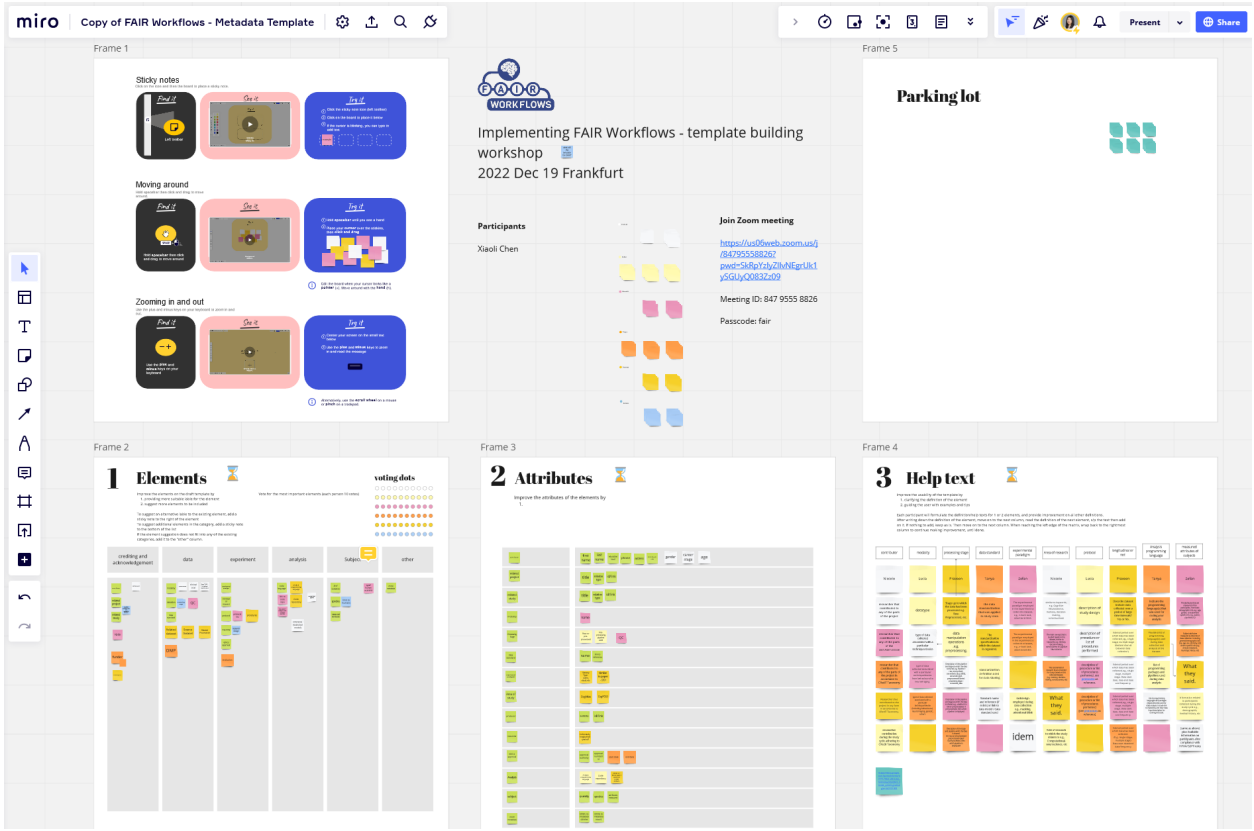


Figure 2. Miro board used in the first template workshop where elements, attributes, and instructional texts were discussed and developed.

The iterated draft template was tested in the workshop with four cognitive neuroscience researchers at the MPI.

Introductory documents were sent to the testers to familiarize them with the project and purpose of the template before the workshop. When the testers arrived at the workshop, they were granted access to the fillable metadata form on the CEDAR platform, and asked to try filling it out and raise any issues they encountered that prevented them from providing suitable information.

The main takeaways from the workshop were:

1. Address specifically the needs of the research workflows of cognitive neuroscience with human subjects;
2. Emphasize metadata to support data discovery (the first step of reuse);
3. Accommodate connection building and FAIR crediting whenever possible.

Date	**Participants**	**Number of participants**
December 16 2022	Project team and guest MPIEA researchers	10
January 26-27 2023	MPIEA researchers	5
February 8-10 2023	Project advisors	5

Table 3. Time and participation of the testing sessions.

The input and feedback collected during the first workshop and testing were implemented before two more rounds of tests were conducted with researchers at MPI and project advisors (Table 3). The template went through iterations following each test based on the feedback gained. In each test, the participants were given either verbal (when the test was done in-person) or written instructions, and the feedback was compiled and addressed systematically. The instruction documents can be found in the appendix.

Template

The OpenView page of the CEDAR template can be accessed via this link. Here we present the first minimum viable product version of the metadata template in a tabular form. The table lists properties (numbers), sub-properties (decimal numbers), and attributes (letters) of the property. The table also lists "occurences" of the properties and attributes, where "1" means required, one value; "0-1" means not required, not repeatable; "0-n" means not required, repeatable; "1-n" means required, repeatable.

The "instruction" column of the table lists textual instructions on the template that helps the user understand the property and the intention of the field. The Ontological constraints column lists the source of ontological terms allowed in the field (when applicable).

ID	Property	Occ.	Instruction	Ontological constraints
1	Dataset	1	Use this section to capture domain-specific information about the dataset, including data modality, acquisition technique, processing stage, domain standards used to arrange and annotate the data, and access information about source data, if available.	
1.a	Data modality	1-n	Select the type of data collected associated with a particular technique/device. Options including behavioral, neuroimaging, genetic, other). Leave the field blank if not applicable.	Controlled list with terms from: National Cancer Institute Thesaurus
1.b	Data acquisition technique	0-n	If the dataset includes neuroimaging data, please indicate the procedure or technology used to collect the data. Leave this field blank if the dataset does not include any neuroimaging data.	Controlled list with terms from: National Cancer Institute Thesaurus
1.1	preprocessing	0-1		
1.1.a	Preprocessing pipeline	0-n	Information about the preprocess used to produce the dataset. Please provide the link to the documentation describing the analysis process, using DOI when possible. (e.g. Brainlife workflow publication)	
1.1.b	Preprocessing script	0-n	Provide a link to the GitHub repository hosting the preprocessing code. To ensure the accessibility and compatibility of the code, consider depositing a copy of the code together with the dataset following the Dryad submission process.	
1.2	Standard	0-n		

1.2.a	Data standard	0-1	Community data standard according to which the dataset is organized.	Controlled list: "Brain Imaging Data Structure (BIDS)" "Neuroimaging Data Model (NIDM)" "Neuroscience information exchange format (NIX)" "Other"
1.2.b	Reference to data standard	0-1	Provide a link to authoritative documentation on the data standard used for the dataset. Use a persistent identifier (i.e. DOI) when possible.	
1.3	Source	0-n		
1.3.a	Source data identifier	0-n	Only relevant when uploading preprocessed or higher-level data - If the source dataset is publicly available, please provide the link here (use persistent identifier, i.e. DOI when possible).	
1.3.b	Source data manager	0-n	If the source dataset is not openly accessible, please provide contact information to the responsible person or team who can provide further information. Leave blank if not applicable.	
1.3.c	ORCID of source data manager	0-n	Please provide the ORCID of the manager of the raw data. Repeat the field to enter multiple persons	
2	Experiment	1-n	This section captures experiment related information, including general area of research, links to clinical trial registration, experimental protocol, ethics approval, as well as research design, experimental paradigm used, and information about participants in the experiment. Please leave the fields blank if the information is unavailable.	

2.a	Area of research	1-n	The central field of research that is intended for study, based on this collected dataset. (e.g. memory, decision-making, consciousness, etc) The field is supported by ontological terms in the Concept category of Cognitive Atlas Ontology. The template will attempt to auto-fill based on the first letters of your input. If no direct matching appears, try to pick one that is closest to your field.	
2.b	Clinical trial	0-n	A clinical trial is a research study that prospectively assigns human participants or groups of humans to one or more health-related interventions to evaluate the effects on health outcomes. If the dataset is produced during a clinical trial experiment, please provide the link to the trial registration. (Use persistent identifier i.e. DOI when possible.) Leave blank if it does not apply.	
2.c	Protocol	0-n	An experimental protocol is a description of procedure or list of procedures performed (see protocol.io as reference). Provide a link to the publicly accessible record of the protocol. Use a persistent identifier i.e. DOI if possible. Leave blank if it does not apply.	
2.d	Preregistration	0-n	Preregistration is a practice of documenting your research plan at the beginning of your study and storing that plan in a read-only public repository such as OSF Registries or the National Library of Medicine's Clinical Trials Registry. Provide a link to the publicly accessible record of the preregistration. Use a persistent identifier i.e. DOI if possible. Leave blank if it does not apply.	

2.1	Ethics approval	0-n		
2.1.a	Ethics approval number		Provide unique number of identifier of the ethics approval	
2.1.b	Start date		Specify the start date of the data being publicly available, based on the ethics document.	
2.1.c	End date		Specify the end date of the data being publicly available, based on the ethics document. If no end date is assigned, leave the field blank.	
2.1.d	Approval authority		Organization that reviewed the project and provided the ethics approval.	
2.1.e	Approval document		If publicly available, provide a link (persistent identifier when possible) to the ethics approval document.	
2.e	Research Design	1-n	The overall strategy used to carry out research. E.g. Longitudinal study design, Block design, etc.	
2.2	Paradigm	0-n		
2.2.a	Experimental paradigm	1-n	The experimental paradigm/ task design employed in the experiment to collect the dataset, e.g., n-back task, attentional blink. The ontological terms accepted in this field are based on the Task category of the Cognitive Atlas Ontology. The template will attempt to auto-complete your entry based on the first letters, please be patient as the system searches for acceptable terms. Please leave the field blank if the paradigm you used is not present in the ontology.	Controlled list with terms from: Cognitive Atlas Ontology
2.2.b	Publication describing paradigm	1-n	Provide the link to the publication or record that describes the paradigm.	
2.2.c	Sensory modality	1-n	The modality of the sensory perception of a stimulus.	Controlled list with terms from: Cognitive Paradigm Ontology
2.3	Subject	0-1		

2.3.a	Number of subjects	0-1		
2.3.b	Demographic attributes measured	0-n	Demographic qualities related to participants collected during the study cycle e.g. age, sex, ethnicity, etc.	Controlled list with terms from: Neuroscience Information Framework (NIF) Standard Ontology (NIFSTD)
2.3.c	Medical attributes measured	0-n	Clinical condition information related to participants collected during the study cycle e.g. autism spectrum disorder, cardiovascular system disease, etc.	Controlled list with terms from: Neuroscience Information Framework (NIF) Standard Ontology (NIFSTD)
3	Analysis	1-n	This section captures information about analysis pipeline and statistical models used on the processed data, as well as the scripting language of the analysis code.	
3.a	Analysis pipeline	0-n	Information about the analysis processing steps used to produce the dataset, or on the dataset. Please provide the link to the documentation describing the analysis process, using a DOI when possible. (e.g. Brainlife workflow publication)	
3.b	Statistic model	0-n	Statistic tests used in data analysis.	Controlled list with terms from: Statistics Ontology
3.c	Programming language	0-n	The language in which source code is written, intended to be executed/run by a software interpreter. Please enter the language of the code that is uploaded along with the dataset.The field is supported by the Software Ontology. If the language you used is not listed, please use the following field (Programming Language (Other)) to indicate the name of the language.	Controlled list with terms from: Software Ontology
3.d	Programming language (other)	0-n		
4	Crediting and Acknowledgement	0-1		

4.1	Contributor	0-n	Researcher that contributed to the project in any form. Enter in addition to the data authors listed on the Dryad data submission form. Repeat this element if multiple contributors are to be included.	
4.1.a	First name	0-1		
4.1.b	Last name	0-1		
4.1.c	ORCID	0-1		
4.1.d	Contributor role	1-n	A high-level classification of the diverse roles performed in the work leading to a published research output in the sciences. Its purpose is to provide transparency in contributions to scholarly published work, to enable improved systems of attribution, credit, and accountability. More on CRediT Taxonomy https://credit.niso.org/ Repeat the field to enter multiple roles for this contributor.	Controlled list with terms from: Contributor Role Ontology
4.1.e	Career stage	0-n	Please indicate the highest degree obtained by the contributor. Leave blank if the contributor is not on an academic career path or prefer not to share the information.	Controlled list with terms from: National Cancer Institute Thesaurus
4.2	Related project	0-n	Information about the project or projects in the course of which this dataset was produced.	
4.2.a	Project title	0-1		
4.2.b	Project identifier	0-1	Provide the link to the public record of the project. Use a persistent identifier (e.g. DOI, RAiD) when possible.Repeat the element (name and identifier) if multiple projects were involved and should be associated with the dataset.	
5	Further information	0-1		

5.a	Link to full metadata	0-n	If other metadata schema was used (e.g. GENEMEDE) during the research process and additional metadata are available for reference, please provide a link to the resource. Use a persistent identifier (e.g. DOI) when possible.	

Table 4. Metadata template for cognitive neuroscience research with human subjects in tabular form.

Challenges

The experience of building the domain-specific template presented several challenges.

Reaching consensus in domain taxonomy and ontology

One of the main challenges that emerged in the process of building a domain-specific metadata template was to identify key properties (or taxonomic categories) to include and what values (ontological terms) can be used as descriptors for each property. Categories usually have fuzzy borders and different terms are sometimes used to refer to the same concept. It is advisable to start small in terms of scope and key properties. More extensive surveys and iterations are required to make a metadata tool robust and usable in the domain-specific context.

Unclear scope and lack of incentives

For the researchers, while a successfully implemented template will provide great value in the long term, there is a steep learning curve and is time-consuming to build. The resources are scattered and difficult to harmonize; for this reason, it is difficult to arrive at a point of consensus in terms of use cases and the overall value of minimal domain data descriptors. There is also a lack of clear incentive for researchers and research teams to adopt one standard over another, or converge on (or agree on) a given metadata schema or profile. While the metadata template is built as a starting point of capturing and surfacing domain-specific metadata to the open, without discovery tools that actually make use of this metadata, it is difficult for the researchers community to commit to the effort of contributing to the infrastructure.

Sustainability model

Ideally a metadata template - effectively a community standard - should be maintained kept up to date long term to ensure effectiveness and usability. A framework for governance and an open channel for feedback and contribution from the community is necessary for this to happen. The

research team considered to adopt a similar approach as the BIDS Extension Proposal process[5] to provide a guide and channels for the community to provide systemic feedback. This approach will include:

1. Suggesting terms to include from accepted ontologies
2. Suggesting terms to be included in specific ontologies

However, with time-bound projects, a sustained maintenance workflow is hard to be followed up on.

Working with an emerging tool

Another challenge stems from the adoption of emerging research tools. The usability of the CEDAR workbench and webforms rendered from templates are pain points both in the process of template building and completing the form. From the template builder's perspective, when a draft template is put together, making small iterative changes on specific fields, such as inserting or editing instructional texts in the form during the testing process is time consuming due to the way nested fields and elements on CEDAR are structured. CEDAR does currently not support responsive form functions (allowing the form to preselect values in fields based on data entered in previous fields), making it difficult to accommodate a wider scope on one template. From the user's perspective, the form is hard to navigate when it comes to repeatable fields, and sometimes loses data due to interface design choices.

Next steps

The immediate next step is to further evaluate the feedback we have received from the community that is not implemented in the Minimum Viable Product version of the template.

Suggested changes and additions under consideration

Through extensive testing, we received a number of suggestions to improve the template based on its current form. Below, a selection of the suggestions are listed in categories.

Suggestion	Field
Dataset	
Provide options for data acquisition techniques based on data modality.	Data acquisition technique
Add descriptive information about data quality, i.e. what type of quality checks are performed.	TBD

[5] https://bids.neuroimaging.io/bep_guide.html

Add information about whether data across different modalities are synchronized.	Data Modality
Request the user to create a release of the version of the code that was used for preprocessing, deposit the code, and ask for the DOI of the published code release.	Preprocessing
Experiment	
Add information about the number of sessions/blocks/trials that were conducted in the experiment.	Paradigm
Adding information about the nature of stimuli (natural or artificial).	TBD
For neural data, add information about anatomical regions of interest.	TBD
Add information about whether data is collected during resting-state or task-state.	TBD
Analysis	
Add information that addresses the questions: What variables were manipulated? How were they manipulated?	TBD
Currently, only "what attributes are measured" is collected, not "how were the attributes measured", possibly to expand this section.	Attributes measured
Adding language options (Julia, awk) beyond the scope of Software Ontology.	Programming language

Table 5. Samples of suggestions from testing that require further evaluation.

Domain ontology based on keyword mapping

As mentioned in the approach section, a project dedicated to ontology building would be desirable follow up to the exploratory efforts made in the beginning of the metadata template building process. With more time and resources the researchers can employ a much more rigorous method to develop a sound taxonomy for the domain (and sub-domain) concepts, as well as a comprehensive and evolving corpus of ontological terms, based on which series of templates can be configured to fit different use cases and cater to various implementation requirements.

Reference

Devaraju, Anusuriya, Huber, Robert, Mokrane, Mustapha, Herterich, Patricia, Cepinskas, Linas, de Vries, Jerry, L'Hours, Herve, Davidson, Joy, & White, A. (2022). FAIRsFAIR Data Object Assessment Metrics. https://doi.org/10.5281/ZENODO.6461229

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Scientific Data, 3(1), 160044. https://doi.org/10.1038/sdata.2016.44

Musen, M. A., O'Connor, M. J., Schultes, E., Martínez-Romero, M., Hardi, J., & Graybeal, J. (2022). Modeling community standards for metadata as templates makes data FAIR. Scientific Data, 9(1), 696. https://doi.org/10.1038/s41597-022-01815-3

Poline, J.-B., Kennedy, D. N., Sommer, F. T., Ascoli, G. A., Van Essen, D. C., Ferguson, A. R., Grethe, J. S., Hawrylycz, M. J., Thompson, P. M., Poldrack, R. A., Ghosh, S. S., Keator, D. B., Athey, T. L., Vogelstein, J. T., Mayberg, H. S., & Martone, M. E. (2022). Is Neuroscience FAIR? A Call for Collaborative Standardisation of Neuroscience Data. Neuroinformatics, 20(2), 507–512. https://doi.org/10.1007/s12021-021-09557-0

Appendix

Instruction for expert feedback

For testing of the draft template project advisors were invited via email to view, fill, and give feedback on the metadata form on the CEDAR platform. They were sent the following information as instructions for their review.

Context

- The exercise is part of the Implementing FAIR Workflows project effort, specifically building metadata templates for cognitive neuroscience research on human subjects.
- For this exercise, we focus on one particular template that will be implemented on the Dryad data repository.
- The purpose of the template is to complement the Dryad dataset (data and potentially accompanying code) submission form with domain-specific metadata, to improve visibility and discoverability of data
- The goal of the exercise is to finalize the draft template.

Instructions

- Register for a CEDAR account (possible to authenticate through ORCID)
- Check the template structure on the CEDAR Open View page
- Fill in the template (login required) with a specific dataset in mind
- Keep track of how much time it takes to complete the form.
- Provide feedback on the following aspects
    - Instruction: How precise and understandable are the instructions for correctly using the template.
    - Description: How accurate and useful are the description for each field for filling the form with desired metadata.
    - Coverage: does the coverage of the information make sense - have we included all the key information needed to locate a specific dataset (improve discoverability of the dataset)
    - Length: time used to complete the form
    - Likelihood of adoption: if the template is used on generic data repositories as a way to provide discipline-specific information, how likely will you use it?
- Any other comment or questions

Background

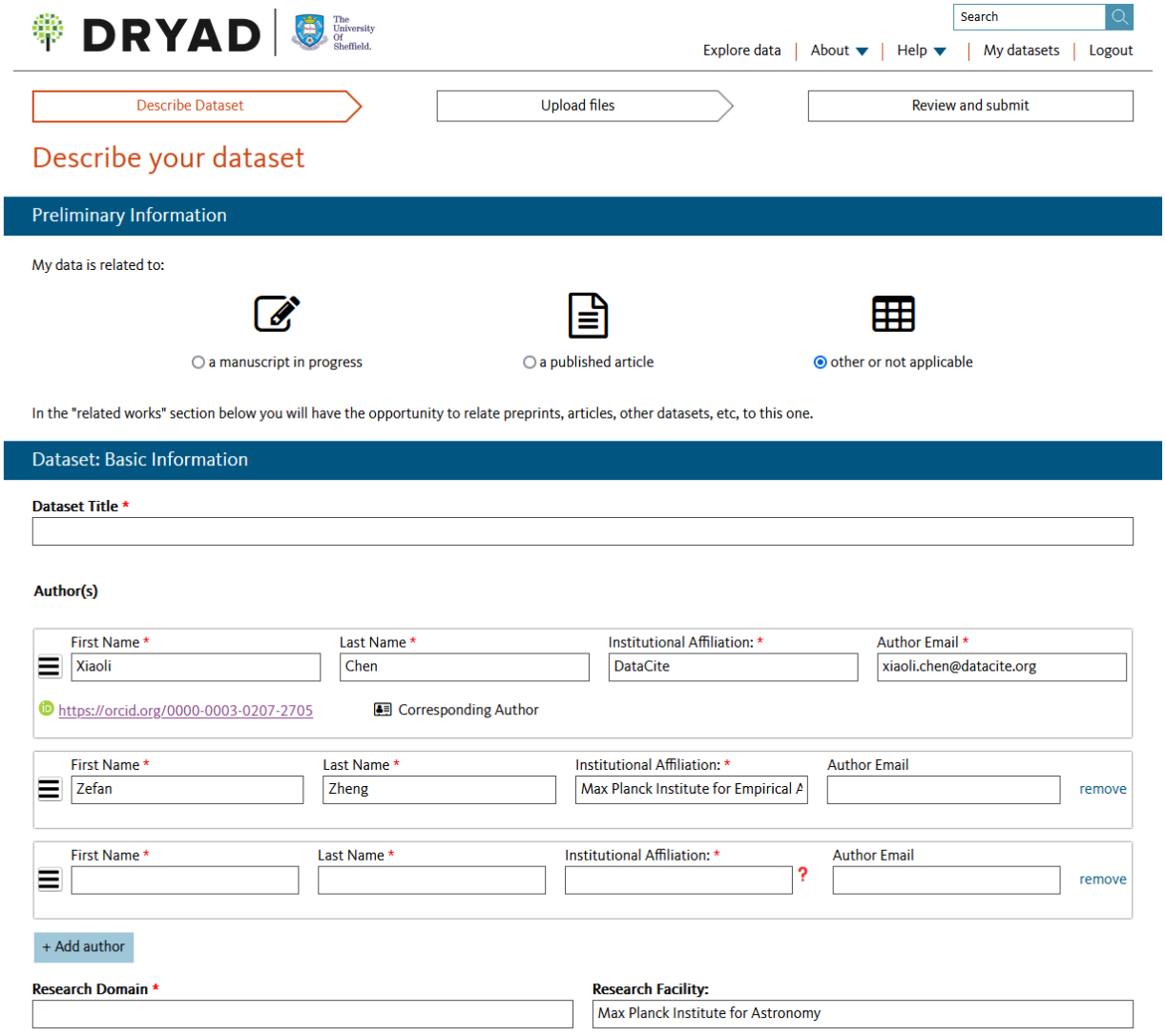
The Implementing FAIR Workflows project

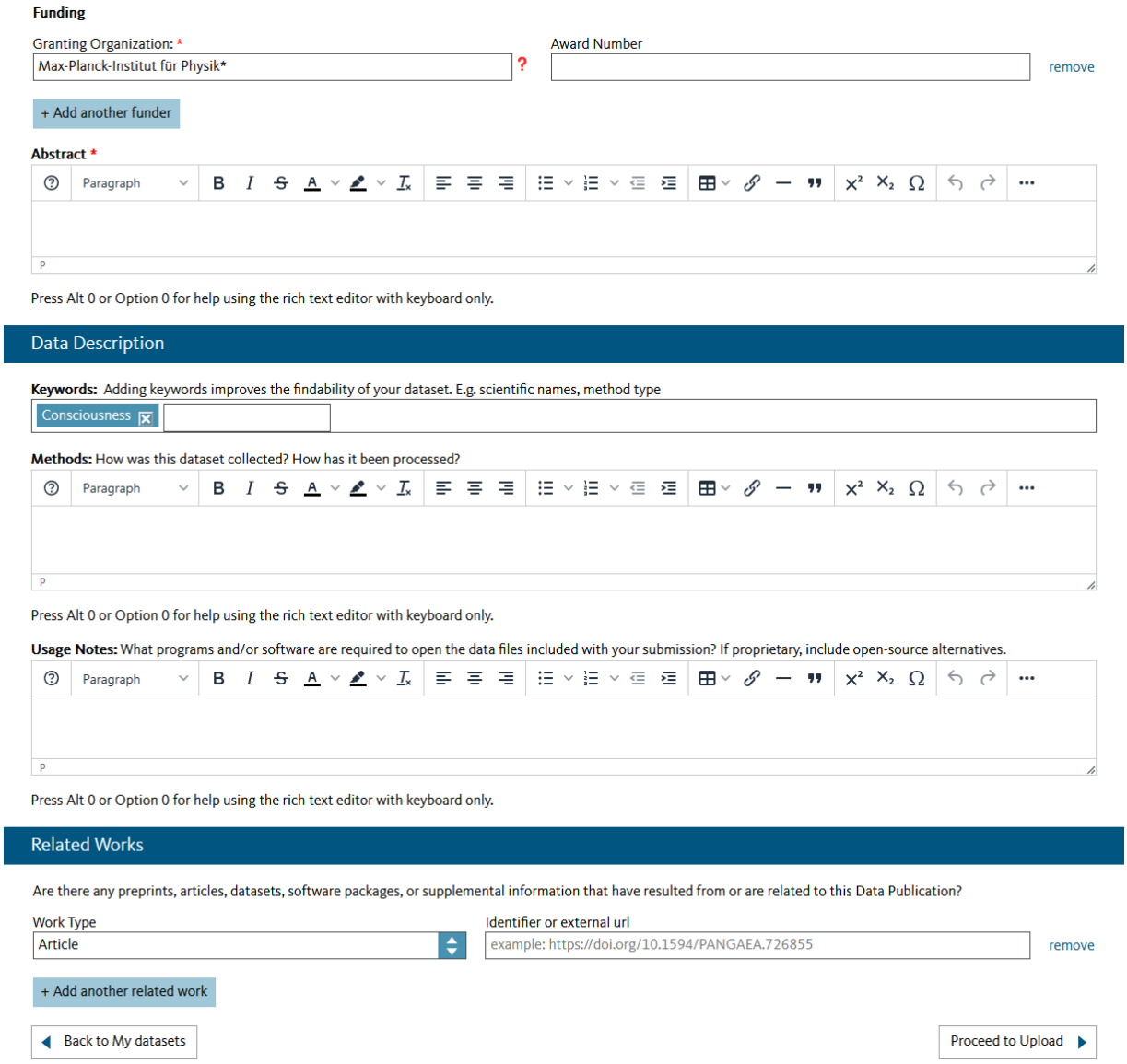
The Implementing FAIR workflows project aims to build an exemplar workflow for researchers and research tools by working with a Cognitive Neuroscience research team to implement FAIR practices throughout a 3 year research project. Leveraging existing services and platforms that are integrated with open persistent identifier (PID) infrastructures, the project charts the way forward for researchers and research teams that are looking to take action towards making their research more reproducible and reusable, as well as potential integrators that are interested to provide support to research workflows and make concrete contributions to the scholarly infrastructure on a metadata level.

In this workshop, we focus on the aspect of effective sharing of research output, specifically dataset from consciousness research. Providing additional metadata that is represented using a formal knowledge representation language, specifies the content of the data, and follows a standard recommended by the target research community significantly increases the FAIRness of the data, thus the formulation of the metadata template.

Dryad

Dryad is a general purpose research data repository. To publish dataset on Dryad, researcher will need to fill in submission form:





The template we built will be implemented as an element on the submission form, specifically a drop down menu will be added to the interface, where the user can select an appropriate template. Once selected, a pop-up window will appear containing a form generated from the template, where the user can fill in relevant information. When finished filling in the form, the user can then save and close the form, and the information generated with the form will be stored in a json file, deposited alongside the dataset that's being published on Dryad.

CEDAR Metadata Template

Check out CEDAR, and this paper to get a better understanding about what metadata templates are all about.

Domain-specific metadata templates built on CEDAR can:

- Help researchers capture/generate rich metadata based on community-based standards
- Provide researchers with the flexibility to be as granular/specific as needed when defining the metadata fields and categories
- Support the template form with semantic resources to make them machine-readable

Excerpts from the project proposal outlying the vision for integrating domain-specific metadata workflows to the research project:

> ... the research group will work with the project partners to prepare templates, work on schemata and ontologies that map to domain specific metadata, and to prepare data and software for open sharing. The group is a leader in the field of Consciousness research and will therefore be able to provide the domain expertise needed to develop such templates in collaboration with CEDAR.
>
> ... We will take advantage of state of the art standards and best practices in the human neuroscience community such as standards for open data sharing i.e,. BIDS for MEG and MRI, will participate in ongoing efforts to extend those standards to Eye tracking, and follow different cognitive ontologies that have been proposed e.g., Cognitive atlas and the cognitive paradigm ontology. Critically, we will extend efforts in the area of description of experimental paradigms including stimuli and tasks, which while key for data aggregation across studies and generalization of the findings and paramount to the discovery of the neuronal correlate of consciousness as detailed below, have received less attention by the community (see NIDM-experiment). The latter is a major challenge, to make progress, and to contribute to the research community studying consciousness, we will focus on developing domain specific descriptions and templates for sharing metadata related to experiments in the field of consciousness to be further extended by our community.
>
> ... Increased discoverability and reuse of the research outputs: The research outputs produced by the research study will be made available with high quality FAIR metadata (using the metadata templates).

CEDAR user scenarios https://more.metadatacenter.org/purpose/scenarios