A standard language for the description of datasets obtained in experimental studies

Alena Begler[1][0000-0003-4375-1106]

[1] St. Petersburg University, St. Petersburg 199034, Russia
alena.begler@gmail.com

Abstract. Despite the number of data increases rapidly this does not lead to a comparable increase in knowledge. This is particularly topical for the data received in scientific research as research efforts are expensive and publicly funded. The fundamental possibility of data reuse is provided by metadata but there are a set of standards that often not consistent. Thus, if two datasets described with different standards it is not easy to integrate them. This paper proposes a language for the description of datasets obtained in behavioral experiments. The language allows connecting datasets obtained by independent research groups. The language consists of two top-level concepts – for the experimental procedure and for the resulting dataset description – and ten lower-level concepts. Each of the concepts is described by mandatory and additional characteristics. The former is necessary for the data description, while the latter improves the understanding of the dataset and its suitability for reuse. The developed language was tested on the description (and further integration) of visual search task datasets obtained by different researchers.

Keywords: metadata, dataset description language, metadata schema, experimental studies datasets.

1 Introduction

Despite the number of data increases rapidly this does not lead to a comparable increase in knowledge. At the same time, sustainability problems are rising – we are not living in the open world anymore and with the modern growth pace, we will sooner or later run out of resources (Weizsäcker von & Wijkman, 2018). This is particularly topical for the scientific research data as research efforts are expensive and publicly funded.

The majority of research data are not reusable: it is neither managed properly as separate datasets (Vines et al., 2014) nor integrated between each other (Wilkinson et al., 2016). Thus, to test any hypotheses researcher should collect his or her own dataset, though some of them could be tested with the datasets already collected by other researchers. This issue can be solved with a joint database, where different datasets could be connected. Current research repositories such as figshare[1], Zenodo[2], or Open

[1] https://figshare.com/

Science Framework[3] do not suite this task – several hundreds of them exist, each one with its own metadata set and store policies.

To connect different datasets universal language for their description should be created. Several attempts were already done; however, neither of them was widely applied. These attempts can be divided into three groups:

1. Ontologies for experimental studies, such as EXPO (Soldatova & King, 2006) and SWRC (Sure, Bloehdorn, Haase, Hartmann, & Oberle, 2005).
2. Domain-specific ontologies for data integration like BrainMap[4] (Fox et al., 2005), NeuroSynth[5] (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011) and Neuroscience Information Framework[6] (Gardner et al., 2008) for neuroscience research.
3. Metadata sets, for example, DCAT (Archer, 2014) and DataCite Metadata Schema (Starr & Gastl, 2011).

In this paper, a universal language for describing datasets of behavioral experiments is proposed. The pilot version of the language was created and tested. Further expansion to the different experimental tasks is under development. After this, the language can be used in two ways: as a metadata schema for internal use in research projects and for the integration of datasets obtained by different research groups. For the last case, a prototype of the platform is planned to be created.

2 Method

To create a data description language, the NeOn approach was adapted (Suarez-Figueroa, Gómez-Pérez, & Fernández-López, 2012). The development included five steps:

4. Specification. The language should meet the requirements:
   a. allows reuse of experimental data;
   b. permits dataset description with a minimum set of required characteristics;
   c. suitable for the data obtained in different behavioral experiments.
5. Analysis of the ontological and non-ontological sources for reuse (briefly presented in the Introduction).
6. Conceptualization. A dictionary of concepts describing datasets was created.
7. Formalization. Hierarchy of the concepts was created (ten upper-level concepts were identified) and their properties were defined (the concepts were divided into two groups, for each of the concept characteristics were formulated).
8. Implementation. YAML was chosen as a human-readable language for describing the data structure. In this language, the data is described using the ‘parameter: value’ pairs, where the parameters are language schema, and the values are added in

[2] https://zenodo.org/

[3] https://osf.io/

[4] http://www.brainmap.org/

[5] http://neurosynth.org/

[6] https://neuinfo.org/

accordance with the dataset. We are aiming to change it to OWL together with graphical user interface implementation.

3 Results

Developed experimental data description language includes two top-level concepts: for data and metadata description (Figure 1)[7]. Metadata (`Meta`) contains information that is necessary for dataset understanding and reuse: authors' names, usage rules, file names, etc. The data description (`Dataset`) includes information about the experimental approach and variable definitions. Metadata descriptions and data descriptions contain `Required` characteristics without which data cannot be used, and `Optional` characteristics that improve the understanding of the dataset and its suitability for reuse. Thus, there are four types of characteristics for the dataset description: 1) metadata required; 2) metadata optional; 3) data required; 4) data optional.

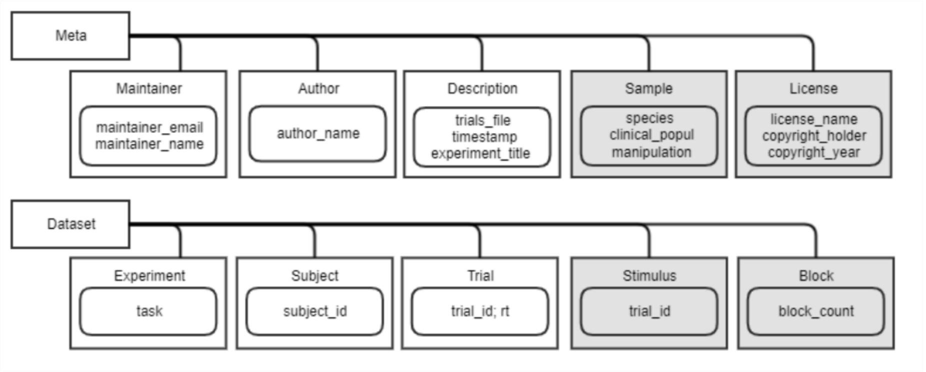


Fig. 1. Schema of the language for datasets description. Top-level elements are in the rectangles, required properties are in the rounded rectangles. Optional characteristics are marked with the grey filling.

Metadata characteristics descriptions vary little from experiment to experiment. It is similar to the common metadata standards such as Dublin Core (Weibel, Kunze, Lagoze, & Wolf, 1998) with two main additions. The first is disambiguation of the experiment's title (`experiment_title`) and the name of the data file (`trial_file`) in the `Description` characteristics as the same experiment can contain several files. The second is `Sample` characteristics as they are crucial for the data reuse.

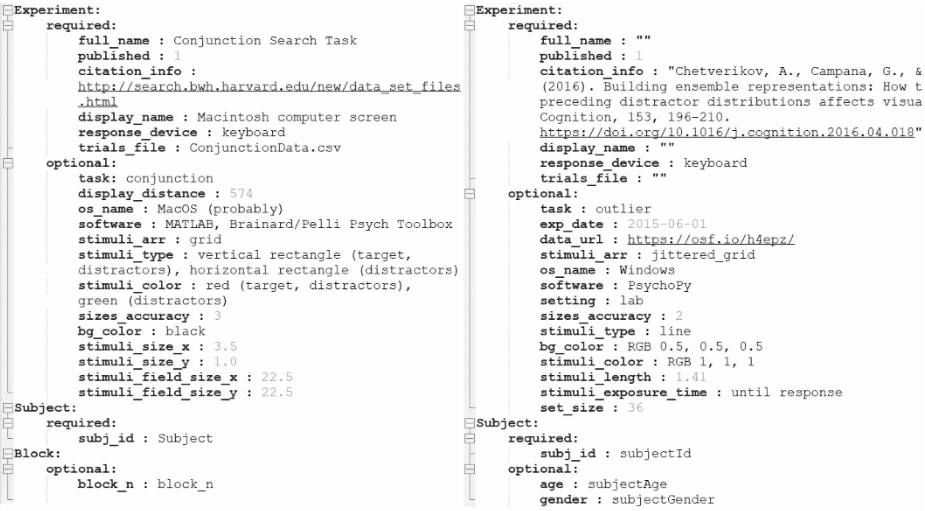
Data characteristics vary significantly in different experiments. For example, in an experimental study of visual search, a set of variables describe stimuli parameter – its spatial (size and location), temporal (for how long it was demonstrated, the interval

[7] Full language schema available freely at https://github.com/achetverikov/visual_search_db/blob/master/data/import_conf_template.yaml

between different stimuli) and other visual (like color or shape) characteristics. In the experiments studying subjective experience, this set will be different – the characteristics related to the stimuli assessment will be in the first place (such as likability scale). To preserve suitability for such different datasets' description and universality of the language simultaneously detailed description of the task was moved to the optional characteristics. Thus, in the developed language, **required data characteristics** are not specific to the task and include a general description of the experiment and participants. **Optional data characteristics** allow to describe the experimental task in more details and include characteristics of experimental conditions (language, date and location, software); equipment (display parameters, characteristics of the response device); and procedure (type and characteristics of the task, parameters of the presented stimuli).

4 Case study

The pilot version of the language was tested out on the datasets obtained in the visual search task. The datasets were collected at open research repositories and websites of the several research groups in the field. The language was able to describe datasets obtained by different authors (Figure 2). Based on the description the datasets were merged into a single database using the R software environment[8] and Neo4j graph database platform[9].



```
Experiment:
    required:
        full_name : Conjunction Search Task
        published : 1
        citation_info :
        http://search.bwh.harvard.edu/new/data_set_files
        .html
        display_name : Macintosh computer screen
        response_device : keyboard
        trials_file : ConjunctionData.csv
    optional:
        task: conjunction
        display_distance : 574
        os_name : MacOS (probably)
        software : MATLAB, Brainard/Pelli Psych Toolbox
        stimuli_arr : grid
        stimuli_type : vertical rectangle (target,
        distractors), horizontal rectangle (distractors)
        stimuli_color : red (target, distractors),
        green (distractors)
        sizes_accuracy : 3
        bg_color : black
        stimuli_size_x : 3.5
        stimuli_size_y : 1.0
        stimuli_field_size_x : 22.5
        stimuli_field_size_y : 22.5
Subject:
    required:
        subj_id : Subject
Block:
    optional:
        block_n : block_n
```

```
Experiment:
    required:
        full_name : ""
        published : 1
        citation_info : "Chetverikov, A., Campana, G., &
        (2016). Building ensemble representations: How t
        preceding distractor distributions affects visua
        Cognition, 153, 196-210.
        https://doi.org/10.1016/j.cognition.2016.04.018"
        display_name : ""
        response_device : keyboard
        trials_file : ""
    optional:
        task : outlier
        exp_date : 2015-06-01
        data_url : https://osf.io/h4epz/
        stimuli_arr : jittered_grid
        os_name : Windows
        software : PsychoPy
        setting : lab
        sizes_accuracy : 2
        stimuli_type : line
        bg_color : RGB 0.5, 0.5, 0.5
        stimuli_color : RGB 1, 1, 1
        stimuli_length : 1.41
        stimuli_exposure_time : until response
        set_size : 36
Subject:
    required:
        subj_id : subjectId
    optional:
        age : subjectAge
        gender : subjectGender
```

Fig. 2. An example of a description of two different datasets obtained by independent research groups. Datasets can be found at http://search.bwh.harvard.edu/new/data_set_files.html и https://osf.io/h4epz/ correspondingly.

[8] https://www.r-project.org/

[9] https://neo4j.com/

Despite the language was able to describe different datasets several **limitations** exist. The main is user-driven enrichment of the language assumed by the current approach for the language extension. It is known that user-tagging systems are redundant and error-prone (Kiu & Tsui, 2011), thus, either the introduction of joint language editing with new characteristics the pre-moderation or its integration with the formal descriptions of experiments is needed. Also, the issue of missing data is not yet taken into consideration.

5 References

Archer, P. (2014). Data Catalog Vocabulary (DCAT). Retrieved July 20, 2019, from https://www.w3.org/TR/vocab-dcat/

Fox, P. T., Laird, A. R., Fox, S. P., Fox, P. M., Uecker, A. M., Crank, M., … Lancaster, J. L. (2005). BrainMap taxonomy of experimental design: Description and evaluation. *Human Brain Mapping*, *25*(1), 185–198. https://doi.org/10.1002/hbm.20141

Gardner, D., Akil, H., Ascoli, G. A., Bowden, D. M., Bug, W., Donohue, D. E., … Williams, R. W. (2008). The Neuroscience Information Framework: A Data and Knowledge Environment for Neuroscience, *6*, 149–160. https://doi.org/10.1007/s12021-008-9024-z

Kiu, C. C., & Tsui, E. (2011). TaxoFolk: A hybrid taxonomy-folksonomy structure for knowledge classification and navigation. *Expert Systems with Applications*, *38*(5), 6049–6058. https://doi.org/10.1016/j.eswa.2010.11.014

Soldatova, L. N., & King, R. D. (2006). An ontology of scientific experiments. *Journal of The Royal Society Interface*, *3*(11), 795–803. https://doi.org/10.1098/rsif.2006.0134

Starr, J., & Gastl, A. (2011). IsCitedBy: A metadata scheme for datacite. https://doi.org/10.1045/january2011-starr

Suarez-Figueroa, M. C., Gómez-Pérez, A., & Fernández-López, M. (2012). The NeOn Methodology for Ontology Engineering. In *Ontology Engineering in a Networked World* (pp. 9–34). https://doi.org/10.1007/978-3-642-24794-1

Sure, Y., Bloehdorn, S., Haase, P., Hartmann, J., & Oberle, D. (2005). The SWRC ontology – Semantic Web for research communities. *Proceedings of the 12th Portuguese Conference on Artificial Intelligence – Progress in Artificial Intelligence (EPIA 2005)*, *3803*, 218–231. https://doi.org/10.1007/11595014_22

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., … Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current Biology*, *24*(1), 94–97. https://doi.org/10.1016/j.cub.2013.11.014

Weibel, S., Kunze, J., Lagoze, C., & Wolf, M. (1998). *Dublin Core Metadata for Resource Discovery*. Retrieved from http://purl.org/metadata/dublin_core

Weizsäcker von, E. U., & Wijkman, A. (2018). *Come On!* Berlin, Germany: Springer

Wilkinson, M. D. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670. https://doi.org/10.1038/nmeth.1635