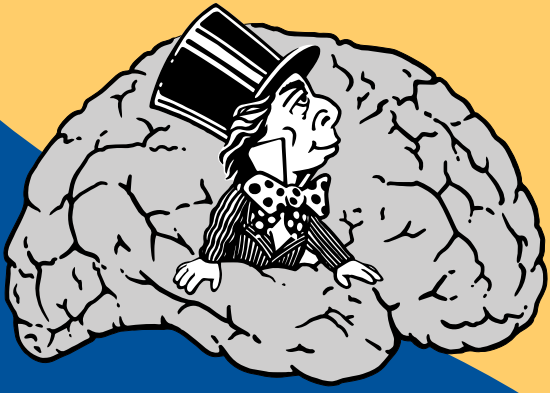
КОГНИТИВНАЯ НАУКА

В МОСКВЕ



НОВЫЕ ИССЛЕДОВАНИЯ

МАТЕРИАЛЫ КОНФЕРЕНЦИИ 2019

Под ред. Е.В. Печенковой, М.В. Фаликман

УДК 159.9
ББК 88.25
К57

Когнитивная наука в Москве: новые исследования. Материалы конференции 19 июня 2019 г. Под ред. Е. В. Печенковой, М. В. Фаликман. – М.: ООО «Буки Веди», ИППиП. 2019 г. — 656 стр.

ISBN 978-5-4465-2346-7

УДК 159.9
ББК 88.25

ISBN 978-5-4465-2346-7

ЯЗЫК ОПИСАНИЯ МАССИВОВ ДАННЫХ ПОВЕДЕНЧЕСКИХ ЭКСПЕРИМЕНТОВ

А. М. Беглер
alena.begler@gmail.com
Высшая школа менеджмента СПбГУ, Санкт-Петербург

Аннотация. Количество получаемых исследователями данных возрастает с каждым годом, что, однако, не приводит к сопоставимому приросту знаний из-за сложности использования этих данных. Принципиальная возможность повторного использования данных обеспечивается наличием метаданных. В данной работе предлагается язык описания массивов данных поведенческих экспериментов. Он состоит из двух понятий верхнего уровня: для описания процедуры эксперимента и для описания полученного в результате массива данных. Каждое из понятий включает по пять понятий нижнего уровня и описывается рядом характеристик, часть из которых — обязательные (без них невозможно воспользоваться данными), а большинство — дополнительные (улучшающие понимание массива и его пригодность для повторного использования). Разработанный язык был применен для описания массивов данных, полученных разными исследователями в задаче зрительного поиска, и их интеграции.

Ключевые слова: метаданные, язык описания массива данных, схема метаданных, интеграция данных, данные поведенческих экспериментов

Введение

Количество получаемых исследователями данных возрастает с каждым годом, что, однако, не приводит к сопоставимому приросту знаний — данные не всегда сопровождаются метаданными, выкладываются на разных репозиториях, в проприетарных форматах и так далее[1]. Принципиальная возможность интеграции данных обеспечивается наличием метаданных, стандартов которых существует несколько десятков[2]. В том числе есть ряд метаданных для экспериментальных исследований: EXPO (Soldatova, King, 2006), онтология научной деятельности (Загорулько и др., 2016), SWRC Ontology (Sure et al., 2005). Обзор для когнитивных исследований можно найти в (Poldrack, Yarkoni, 2016). Интеграция данных исследований с помощью подобных схем распространена

1 Представление о проблемах открытых научных данных дают принципы FAIR, разработанные для обеспечения возможностей повторного использования данных: https://www.force11.org/group/fairgroup/fairprinciples.

2 Например, список стандартов метаданных Digital Curation Centre включает порядка сорока: http://www.dcc.ac.uk/resources/metadata-standards/list.

в основном в области нейронаук (напр., Dickson et al., 2001; Fox et al., 2005; Yarkoni et al., 2011).

В данной работе предлагается язык описания массивов данных поведенческих экспериментов. Разработанный язык описания массивов данных может использоваться двояко. Во-первых, в качестве схемы метаданных для внутреннего пользования в исследовательских проектах. Во-вторых, для описания массивов, полученных разными исследовательскими группами, и их последующей интеграции.

Метод

Для создания языка описания данных был адаптирован подход NeOn (Suarez-Figueroa et al., 2012). Разработка включала пять шагов.

1. Формулировка требований:
   - должен позволять повторное использование экспериментальных данных;
   - должен давать возможность описать массив как минимальными средствами, так и максимально подробно;
   - должен позволять описывать данные, полученные в разных экспериментальных задачах.
2. Анализ аналогичных разработок (кратко представлен во Введении).
3. Формирование словаря понятий, необходимых для описания массива данных (было выделено десять ключевых понятий).
4. Формирование иерархии понятий и их свойств (понятия были разделены на две группы, для каждого из понятий были сформулированы характеристики, необходимые для его описания).
5. Реализация в формальном языке. Был выбран YAML[3] — человеко-читаемый язык для описания структуры данных. В этом языке данные описываются с помощью пар «параметр: значение», где параметрами выступила разработанная схема языка, а значения добавляются в соответствии с описываемым массивом данных.

Результаты

Разработанный язык описания экспериментальных данных включает два понятия верхнего уровня: для описания данных и метаданных (рис. 1). Метаданные (Meta) содержат информацию, которая необходима для того, чтобы воспользоваться массивом данных: имена авторов, правила использования, названия файлов массива и т. д. Описание массива (Dataset) включает в себя информацию об экспериментальной задаче и описания переменных. Описания метаданных и описания массива содержат обязательные (Required) характеристики, без которых невозможно воспользоваться данными, и дополнительные (Optional), улучшающие понимание массива и его пригодность для повторного использования.

3 https://yaml.org/.

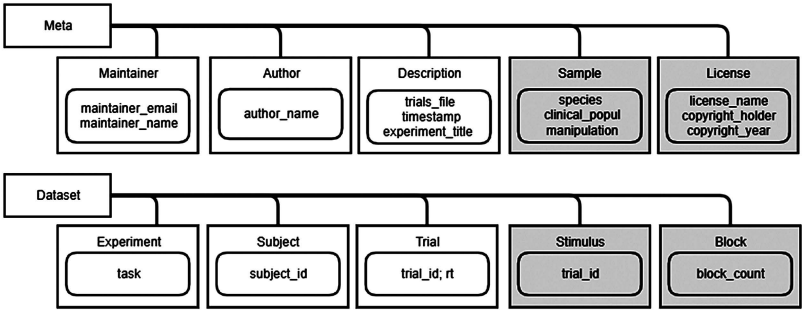


Рисунок 1. Схема языка описания массивов данных. В прямоугольниках — элементы верхнего уровня для описания метаданных (Meta) и данных массива (Dataset). В прямоугольниках с закругленными углами — обязательные свойства каждого из элементов. Серой заливкой помечены элементы, которые могут быть опущены при описании

Если описания метаданных мало варьируются от задачи к задаче, то описания массива могут сильно отличаться для разных задач (сравним, например, экспериментальное исследование зрительного поиска и социальной перцепции). В разработанном языке обязательные характеристики не специфичны для задачи, а дополнительные могут включать как общие для разных экспериментальных задач, так и специфические для конкретной задачи характеристики.

Дополнительные характеристики массива данных включают описание:

- ситуации проведения эксперимента (язык, дата и место проведения, программное обеспечение);
- оборудования (марка и размеры дисплея, характеристики устройства ввода ответа, тип ответа);
- процедуры (тип и характеристики задачи, параметры предъявляемых стимулов).

Описанный язык был применен для описания массивов данных, полученных в задаче зрительного поиска (рис. 2)[4]. Данные описания были использованы для импорта массивов, полученных разными исследователями, в единую базу данных[5].

Обсуждение и выводы

Наибольший интерес представляет использование языка для описания и последующей интеграции массивов данных, полученных разными исследователями. В отличие от других имеющихся разработок, данный язык достаточно гибок: он не привязан к отдельному репозиторию или проекту и может быть изменен в соответствии с текущими исследовательскими задачами и использован для объединения массивов данных, значительно отличающих-

4 Полная схема: https://github.com/achetverikov/visual_search_db/blob/master/data/import_conf_template.yaml.

5 Код выложен на https://github.com/achetverikov/visual_search_db.

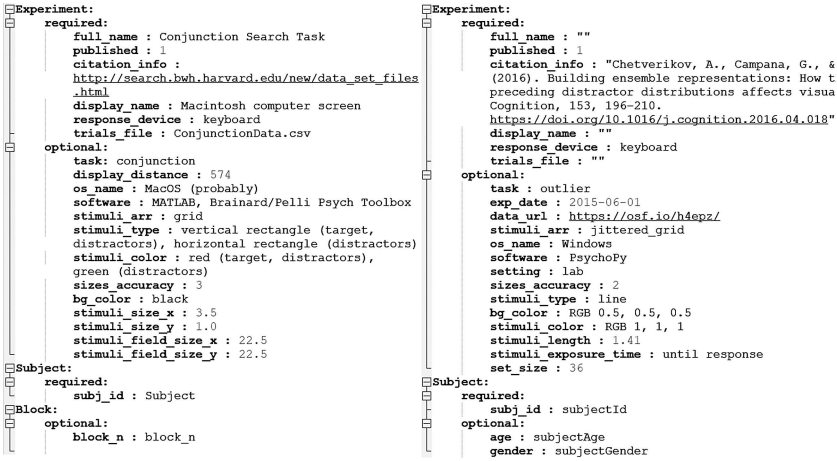


Рисунок 2. Пример описания данных двух разных экспериментов с помощью предложенного языка. Источники данных: http://search.bwh.harvard.edu/new/data_set_files.html и https://osf.io/h4epz/ соответственно

ся по структуре. Однако сложность применения языка для этой цели состоит в том, что ряд характеристик специфичен для разных экспериментальных задач: если наличие маски и фиксационного креста относительно универсально, то, например, статистические характеристики распределения дистракторов важны только для описания задачи зрительного поиска. В текущей реализации данная проблема решена расширяемостью языка, то есть возможностью добавления новых свойств при описании массива. Использованное решение не может считаться окончательным — известно, что системы меток, наполняемые пользователем, страдают от неточности и избыточности (Kui, Tsui, 2011). В дальнейшем необходимо либо введение совместного редактирования языка с премодерацией добавляемых характеристик, либо его интеграция с формальными описаниями экспериментальных задач.

Литература

Загорулько Ю. А., Загорулько Г. Б., Боровикова О. И. Технология создания тематических интеллектуальных научных интернет-ресурсов, базирующаяся на онтологии // Программная инженерия. 2016. Т. 7. №2. С. 21 – 60.

Dickson J., Drury H., Van Essen D. C. 'The surface management system' (SuMS) database: a surface – based database to aid cortical surface reconstruction, visualization and analysis // Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences. 2001. Vol. 356. No. 1412. P. 1277 – 1292. http://dx.doi.org/10.1098/rstb.2001.0913

Fox P. T., Laird A. R., Fox S. P., Fox P. M., Uecker A. M., Crank M., Koenig S. F., Lancaster J. L. BrainMap taxonomy of experimental design: Description and evaluation // Human Brain Mapping. 2005. Vol. 25. No. 1. P. 185 – 198. http://dx.doi.org/10.1002/hbm.20141

Kiu C. C., Tsui E. TaxoFolk: A hybrid taxonomy – folksonomy structure for knowledge classification and navigation // Expert Systems with Applications. 2011. Vol. 38. No. 5. P. 6049 – 6058. http://dx.doi.org/10.1016/j.eswa.2010.11.014

Poldrack R. A., Yarkoni T. From brain maps to cognitive ontologies: Informatics and the search for mental structure // Annual Review of Psychology. 2016. Vol. 67. No. 1. P. 587 – 612. http://dx.doi.org/10.1146/annurev-psych-122414-033729

Soldatova L. N., King R. D. An ontology of scientific experiments // Journal of The Royal Society Interface. 2006. Vol. 3. No. 11. P. 795 – 803. http://dx.doi.org/10.1098/rsif.2006.0134

Suarez-Figueroa M. C., Gómez-Pérez A., Fernández-López M. The NeOn methodology for ontology engineering // Ontology Engineering in a Networked World / M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, A. Gangemi (Eds.). 2012. P. 9 – 34.

Sure Y., Bloehdorn S., Haase P., Hartmann J., Oberle D. The SWRC ontology — Semantic Web for research communities // Progress in Artificial Intelligence. Progress in Artificial Intelligence. EPIA 2005. Lecture Notes in Computer Science Springer Berlin Heidelberg, 2005. Vol. 3803. P. 218 – 231. http://dx.doi.org/10.1007/11595014_22

Yarkoni T., Poldrack R. A., Nichols T. E., Van Essen D. C., Wager T. D. Large-scale automated synthesis of human functional neuroimaging data // Nature Methods. 2011. Vol. 8. No. 8. P. 665 – 670. http://dx.doi.org/10.1038/nmeth.1635

A STANDARD LANGUAGE FOR THE DESCRIPTION OF EXPERIMENTAL DATASETS

A. Begler
alena.begler@gmail.com
Graduate School of Management SPbU, St. Petersburg

Abstract. The amount of data obtained by researchers increases every year, but this does not lead to a comparable increase in knowledge due to the complexity of using this data. The fundamental possibility of data reuse is provided by metadata availability. This paper proposes a language for the description of datasets obtained in behavioral experiments. 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 a set of 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 used for description (and further integration) of visual search task datasets obtained by different researchers.

Keywords: metadata, dataset description language, metadata schema, data integration, behavioral experiment datasets