The best of both worlds: using semantic web with JSON-LD. An example with NIDM-Results & Datalad

Camille Maumet, Satrajit Ghosh, Yaroslav O. Halchenko, Dorota Jarecka,

Nolan B. Nichols, Jean-Baptiste Poline, Michael Hanke

The best of both worlds: using semantic web with JSON-LD. An example with NIDM-Results & Datalad

1. Camille Maumet - Inria, Univ Rennes, Inserm, CNRS, Rennes, France
2. Satrajit Ghosh - Massachusetts Institute of Technology, Cambridge, MA, USA
3. Yaroslav O. Halchenko - Dartmouth College, NH, USA
4. Dorota Jarecka - Massachusetts Institute of Technology, Cambridge, MA, USA
5. Nolan Nichols - Bioinformatics and Computational Biology, Genentech Inc, CA, USA
6. Jean-Baptiste Poline - McGill University, Montreal, Canada, and UC Berkeley, CA, USA
7. Michael Hanke - Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany and Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Introduction

NIDM-Results provides a harmonised representation for fMRI results reporting using Semantic Web technologies (Maumet et al. 2016). While those technologies are particularly well suited for aggregation across complex datasets, using them can be costly in terms of initial development time to generate and read the corresponding serialisations. While the technology is machine accessible, it can be difficult to comprehend by humans. This hinders adoption by scientific communities and by software developers used to more-lightweight data-exchange formats, such as JSON. JSON-LD: a JSON representation for semantic graphs ("JSON-LD 1.1" n.d.) was created to address this limitation and recent extensions to the specification allow creating JSON-LD documents that are structured more similar to simple JSON. This representation is simultaneously readable by a large number of JSON-based applications and by Semantic Web tools. Here we review our work on building a JSON-LD representation for NIDM-Results data and exposing it to Datalad (Halchenko et al. 2018), a data-management tool suitable for neuroimaging datasets with built-in support for metadata extraction and search.

Methods

This work was initiated at two collaborative events: the "INCF Neuroinformatics Hackathon 2018" August 7-8 in Montreal, Canada and "DC Code Convergence" September 10-14 in Washington, DC, United States.

Using JSON-LD 1.1, we developed a context file that can be used to transform the original semantic web file into a simple JSON representation while preserving the PROV data model. "Type indexing", a JSON-LD 1.1 feature, enables indexing objects by their types instead of their identifiers. Fig. 1 provides an example of JSON-LD export on a simplified NIDM-Results graph.
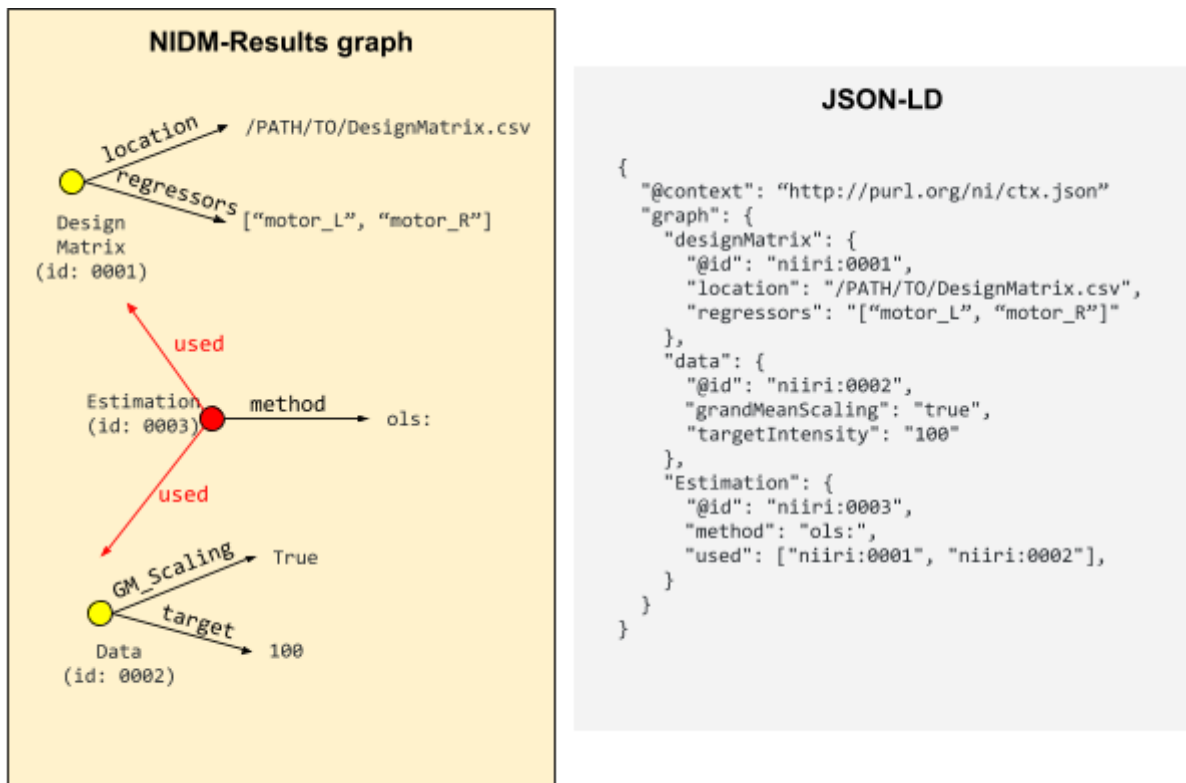
Fig. 1. Example of JSON-LD file using type indexing on a simplified NIDM-Results graph made of 3 objects: "design matrix", "data" and "estimation".

Results

The JSON-LD context file is available at: http://purl.org/nidash/context. The metadata extractor for Datalad is available at: https://github.com/datalad/datalad-neuroimaging/pull/41. Finally, a set of examples of JSON-LD files are included at: https://github.com/incf-nidash/nidm-specs/pull/479. Fig. 2 presents an exemplary invocation and result of a DataLad search, demonstrating that the proposed metadata representation is suitable for processing with simplest text matching procedures without any programmatic "understanding" of term semantics or the need to construct a true graph representation. Accessible metadata representations maximize utility across use cases. Here, DataLad automatically binds NIDM metadata to precise identity and version information, as well as the exact state of any other content of a dataset.

```
% datalad -f json_pp search \
    nidmresults.records.Agent:SPM \
    nidmresults.records.SoftwareAgent.softwareVersion:12b.5858 \
    nidmresults.records.ContrastMap:'passive listen'
{
  "action": "search",
  "dsid": "7a55f7a4-c237-11e8-b964-f0d5bf7b5561",    Dataset ID
  "metadata": {

        not shown: metadata record with all details and full JSON-LD context

    "annex": {
      "key": "MD5E-s4673294--db6067919a0f55f559b91915f2e9b25b.zip"    File content ID
    },
  },
  "parentds": "/tmp/datalad_temp_test_nidm6yx7fm2d",
  "path": "/tmp/datalad_temp_test_nidm6yx7fm2d/mystudy.nidm.zip",
  "query_matched": {
    direct label match, term property match
    "nidmresults.records.Agent.rdfs-label": "SPM",
    partial label match, term property match
    "nidmresults.records.ContrastMap.nidm-NIDM_0000085": "passive listening",
    "nidmresults.records.ContrastMap.rdfs-label": "passive listening",
    direct label match, direct term match
    "nidmresults.records.SoftwareAgent.softwareVersion": "12b.5858"
  },
  "refcommit": "c8ff31326278b65052fbc1f7223d6ba468ce0e49",    Dataset version
  "type": "file"
}
...
```

Fig. 2: Invocation and result of a DataLad search, demonstrating that the proposed metadata representation is suitable for processing with simplest text matching procedures.

Conclusions

Here, we have proposed a new serialisation for NIDM-Results using JSON-LD that is both easier to read by standard applications and equivalent to the original data model, preserving provenance, the graph structure, and merge and query capabilities. Key property of the developed metadata representation is that a large fraction of the contained information is programmatically accessible by direct access of a shallow dictionary-like key store. For any metadata use case where relevant term labels are known, this enables straightforward and complete utilization through generic applications like DataLad, without the otherwise unavoidable technical and conceptual overhead of semantic graphs.

Acknowledgments

References

Halchenko, Yaroslav, Michael Hanke, Benjamin Poldrack, Kyle Meyer, Debanjum, Gergana Alteva, Jason Gors, et al. 2018. *Datalad/datalad 0.11.0*. https://doi.org/10.5281/zenodo.1470735.

"JSON-LD 1.1." n.d. Accessed December 14, 2018. https://json-ld.org/spec/latest/json-ld/.

Maumet, Camille, Tibor Auer, Alexander Bowring, Gang Chen, Samir Das, Guillaume Flandin, Satrajit Ghosh, et al. 2016. "Sharing Brain Mapping Statistical Results with the Neuroimaging Data Model." *Scientific Data* 3 (160102). https://doi.org/10.1038/sdata.2016.102.