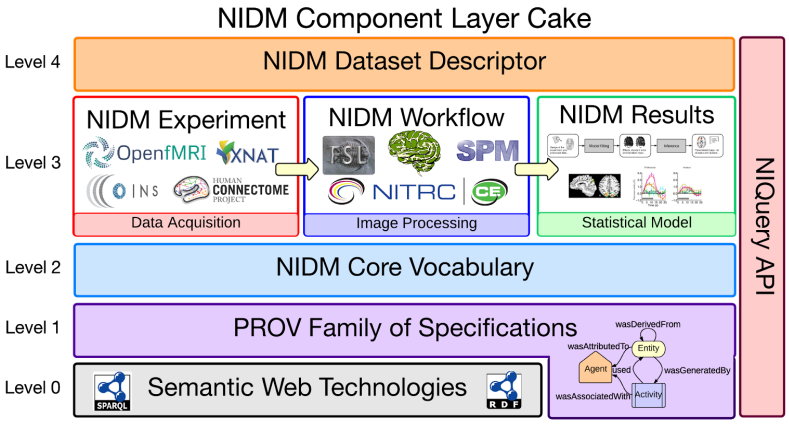
Comments of the use of PROV as the underlying data model for brain imaging (http://nidm.nidash.org)

Building domain models on top of PROV-DM



The NeuroImaging Data Model (NIDM) is being constructed using the PROV Family of Specifications as the basis. Our group is developing object models and a domain specific core vocabulary to capture key elements of human brain imaging including details of the experiment setup, analysis workflow, and results. The vocabulary building involves generating new terms and definitions, and borrowing others when available (e.g., STATO). We incorporate these terms into an object model that captures the dataflow and description of results in brain imaging and thereby represents knowledge of variations across experiments and software platforms. We describe below which tools and guidance were missed in this process.

Use of RDF as core data format for PROV and conversion between PROV-DM and PROV-O

We have committed to using RDF as the base data format, and are therefore using serializations of PROV-DM using PROV-O. We have recently added support for PROV-O in the Python PROV toolbox. One issue is that the Qualification Pattern can be verbose and hard to explain to people for assigning Roles, Plans, etc. to an edge. We have found that our domain experts tend to avoid using these constructs when creating domain models (at least with NIDM Results [1]), and ask questions such as: "How do we assign a role without creating an activity explicitly?" In addition, using the current standard leads to roundtrip failures when converting to PROV-O and back to PROV-N. As Luc Moreau noted (personal communication), "What PROV is missing is a normal form, defined independently of validity. Round tripping should result in documents with the same normal form."

Contention between linked entity identifiers, qualified relations, and SPARQL queries in the PROV graph

The Qualified Relations arise when constructing with PROV-DM and converting to PROV-O. These qualifiers tend to make resulting queries complex. As an example, consider the following pattern in a typical dataflow system; a directed acyclic graph comprising activities that use and generate entities. Using PROV to represent these entities and activities is natural. When exported as RDF, a SPARQL query with property paths should be able to trace paths through these activities, linking source entities to a final activity. Such a query becomes complicated with qualified relations, specifically when most activities we represent use qualified relations when annotated with a 'plan' or 'role' or other attributes. This can be simplified by using 'used' or 'wasGeneratedBy' even for qualified relations, as is reflected in the prov-o document, without resorting to reasoners.

Need for RDF-based tools supporting extended, scalable, and robust visualization and query capabilities

We are very appreciative of the current tools and services for PROV-documents in Java and Python. The visualization tools (such as the Sankey graph) are often not scalable when it comes to practical scientific dataflows consisting of hundreds of activity nodes and many entities. Most domain scientists do not want to write queries in SPARQL so they are reliant on a small subset of developers and informaticians to generate platforms for interacting with the PROV documents. This limits the adoption across our domain. It is important that PROV is usable not just from a data capture perspective, but also from a query perspective. Currently the easiest form for queries is using SPARQL. The toolchains supporting PROV generation as RDF are limited to Java and Python, with even less support for query. Therefore, a coordinated community effort to build tools for common languages is needed.

[1] C. Maumet, T. Auer, A. Bowring, G. Chen, S. Das, G. Flandin, S. Ghosh, T. Glatard, K. J. Gorgolewski, K. G. Helmer, M. Jenkinson, D. Keator, B. N. Nichols, J.-B. Poline, R. Reynolds, V. Sochat, J. Turner, and T. E. Nichols, "NIDM−Results: a Neuroimaging Data Model to share brain mapping statistical results," 2016, doi: 10.1101/041798.