A lightweight approach to research object data packaging

Eoghan Ó Carragáin, Carole Goble, Peter Sefton, Stian Soiland-Reyes

A Research Object (RO) provides a machine-readable mechanism to communicate the diverse set of digital and real-world resources that contribute to an item of research. The aim of an RO is to replace traditional academic publications of static PDFs, to rather provide a complete and structured archive of the items (such as people, organisations, funding, equipment, software etc) that contributed to the research outcome, including their identifiers, provenance, relations and annotations. This is increasingly important as researchers now rely heavily on computational analysis, yet we are facing a reproducibility crisis [1] as key components are often not sufficiently tracked, archived or reported.

We propose Research Object Lite (or RO-Lite for short), an emerging lightweight approach to package research data with their structured metadata, based on schema.org annotations in a formalized JSON-LD format that can be used independent of infrastructure to encourage FAIR sharing of reproducible datasets and analytical methods.

Background

Earlier work introduced the notion of Research Objects [2]. Their formalization combines existing Linked Data standards: W3C RDF, JSON-LD, OAI-ORE, W3C Web Annotations, PROV, Dublin Core Terms, ORCID. The RO ontologies [3] combined these to describe ROs, but do not themselves formalize how ROs are saved or transmitted. Multiple formats have since been realized: the portal RO Hub [4] use RDF REST resources; while workflow provenance make RO bundle ZIP files [5] or Big Data BagIt archives [6, 7]. Each of these require RO support in the packaging infrastructure.

Multiple data packaging initiatives have recently emerged, within Research Data Alliance, Force11, DataOne and elsewhere; like Frictionless data [8] for table-like files, BioCompute Objects for regulatory science [9], CodeMeta for software, Psych-DS for psychology studies, and DataCrate [10] for datasets. RDA has surveyed a large variety of data packaging formats across different domains.

Common among these is structured metadata, e.g. with a single JSON file that refer to neighbouring data files and scripts maintained and published together, e.g. in GitHub. Many of these initiatives use schema.org [11] as basis for common metadata. With JSON-LD this offers a developer-friendly experience and interoperability with web conventions outside of the research domain.

Data packaging principles

At a RDA meeting on data packaging we concluded that many initiatives arrive at similar principles: simple folder structure; JSON-LD manifest; schema.org for core metadata; BagIt for fixity; OAI-ORE for aggregation. This points to: a) appetite for general package/folder-oriented approach in different contexts; b) a generic solution won’t work for all and needs to be domain-extensible; c) a tendency to re-invent the wheel, leading to sub-optimal interoperability and duplication of effort.

Part of this work has been done at the BioExcel CoE (www.bioexcel.eu), a project funded by the European Union contracts H2020-INFRAEDI-02-2018-823830, H2020-EINFRA-2015-1-675728.
We have identified a gap for a solid base format for data packaging that also allow communities to build domain-specific solutions. Frictionless data [8] could arguably fill this gap, with mature specifications and a strong design philosophy, however as an independent JSON format it does not fully apply Linked Data principles, and would be harder to use in FAIR integrations and extensions.

Our evolving proposal, RO-Lite, is based around these principles: a) metadata as Linked Data, using schema.org as much as possible; b) extensible for different domains; c) retain the core Research Object principles Identity, Aggregation, Annotation; d) inferred metadata rather than repetition; e) “just-enough” provenance; f) layered validation; g) archivable with BagIt; h) hooks to reuse existing domain formats; i) lightweight programmatic generation and consumption. Similar to the approach of BioSchemas, rather than building new specifications from scratch, we aim to build best-practice guides and validatable profiles for building rich research data packages with existing standards, without requiring expert knowledge for developing producers and consumers.

Building community consensus

RO-Lite is a fresh initiative, bringing together data archive and repository maintainers with existing Research Object, workflow and provenance communities. Starting as a small cross-domain group, organically formed to build the core principles and first sketches of their use, we are now expanding to collect use cases and reaching out to other packaging initiatives to build common ground.

One emerging use of RO Lite is for capturing workflows and tools in a federated workflow repository being built in EOSC-Life, a large European Open Science Cloud project across 13 research infrastructures in the life science domain. However RO Lite is also aiming to be usable by individual scientists with no particular infrastructure beyond Jupyter notebook, who may not have the time or motivation to use a cascade of metadata vocabularies and research data management tools [12].

RO Lite development and discussion is done openly in a GitHub repository by volunteers, with monthly telcos to synchronize the effort. Anyone can join to help form the RO Lite approach.

References


https://doi.org/10.5281/zenodo.2636997
This work is licensed under a Creative Commons Attribution 4.0 International License.