RO-Crate, a lightweight approach to Research Object data packaging

http://www.researchobject.org/ro-crate/

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 are developing *Research Object Crate* (or **RO-Crate** for short), a 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 (*BDBag*) [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.

Cite as: https://doi.org/10.5281/zenodo.3337883

Data packaging principles

At an 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.

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 proposal is to build on DataCrate [10] to evolve **RO-Crate**, 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.

Challenges

In the desire to make a lightweight approach that is also generally applicable we have focused primarily on RO-Crate as a way to capture and describe a dataset of neighbouring files, for instance maintained in a GitHub repository, while also describing external resources and provenance.

A goal of RO-Crate is to simplify the creation and maintenance of metadata, supporting both a manual or programmatic manner, with a set of opinionated profiles for describing common resource types such as tables, organizations, equipment, software, workflows, places. One design challenge is the pressure between explicit and implicit. For instance, should the creator be repeated for every file they also made, or can the creator of the research object be a reasonable default? Do all files in a directory need to be listed in the RO-Crate metadata, or is it sufficient to describe the directory or a filename glob pattern?

Recognising this challenge, our vision is a separation of concern between *packaging* and *description*. By letting `ro-crate-metadata.jsonld` be used in any file-or web-based approach at a root directory, we can rely on BDBag [7] when the RO-Crate needs to be shipped, archived or deposited. At this point the Research Object would be "completed" to generate a full "classical" Research Object manifest, integrating the BagIt checksums and file list with metadata implied from the RO-Crate metadata. We are developing this tooling for this completion step, which should also perform *validation* against the declared profile, might include snapshotting remote resources and minting of RO identifiers using MinId [7]. Although an RO-Crate is in a sense a recipe for making full Research

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,

Objects, RO-Crate metadata can also be consumed "as-is" as JSON-LD if discovered in the wild, or updated programmatically based on the profile and other schema.org markup.

Building community consensus

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

References

[1] Monya Baker (2016): **1,500 scientists lift the lid on reproducibility**. *Nature* **533**. https://doi.org/10.1038/533452a

[2] Sean Bechhofer et al (2013): **Why Linked Data is Not Enough for Scientists**, *Future Generation Computer Systems* **29**(2) https://doi.org/10.1016/j.future.2011.08.004

[3] Khalid Belhajjame et al (2015): **Using a suite of ontologies for preserving workflow-centric research objects**. *Web Semantics*: Science, Services and Agents on the World Wide Web, https://doi.org/10.1016/j.websem.2015.01.003

[4] Jose Manuel Gomez-Perez et al (2017): **Towards a Human-Machine Scientific Partnership Based on Semantically Rich Research Objects**. *IEEE 13th International Conference on e-Science (e-Science 2017)*. https://doi.org/10.1109/eScience.2017.40 [preprint available]

[5] Stian Soiland-Reyes, Pinar Alper, Carole Goble (2016): **Tracking workflow execution with TavernaProv**. At *ProvenanceWeek 2016; PROV: Three Years Later.* 6 Jun 2016, Washington DC, US. https://doi.org/10.5281/zenodo.51314

[6] Ravi K Madduri et al (2019): **Reproducible big data science: A case study in continuous FAIRness**. *PLoS ONE* **14**(4) https://doi.org/10.1371/journal.pone.0213013

[7] Farah Zaib Khan et al (2018): **Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv**. Accepted for *GigaScience*. (preprint https://doi.org/10.5281/zenodo.3336124)

[8] Jo Barratt, Serah Rono (2018): **Frictionless Data and Data Packages** At *Workshop on Research Objects (RO 2018)*, 29 Oct 2018, Amsterdam, Netherlands. https://doi.org/10.5281/zenodo.1301152

[9] Gil Alterovitz et al (2018): **Enabling Precision Medicine via standard communication of NGS provenance, analysis, and results**. *PLOS Biology*. **16**(12):e3000099 https://doi.org/10.1371/journal.pbio.3000099

[10] Peter Sefton et al (2018): **DataCrate: a method of packaging, distributing, displaying and archiving Research Objects** At *Workshop on Research Objects (RO 2018)*, 29 Oct 2018, Amsterdam, Netherlands. https://doi.org/10.5281/zenodo.1445817

[11] R. V. Guha, Dan Brickley, Steve Macbeth (2016): **Schema.org: evolution of structured data on the web**. *Communications of the ACM* **59**(2). https://doi.org/10.1145/2844544

[12] Cameron Neylon (2017): **As a researcher…I'm a bit bloody fed up with Data Management.** Blog *Science in the Open.* http://cameronneylon.net/blog/as-a-researcher-im-a-bit-bloody-fed-up-with-data-management/ [archived 2019-04-12]

An earlier version of this abstract was accepted for BOSC 2019, see https://doi.org/10.5281/zenodo.3250687

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.