Introducing RO-Crate: research object data packaging

Presenter: Peter Sefton

Peter Sefton[1], Eoghan Ó Carragáin[2], Carole Goble[3], Stian Soiland-Reyes[3]

[1] University of Technology Sydney, Sydney, Australia, Peter.Sefton@uts.edu.au

[2] University College Cork, Cork, Ireland, eoghan.ocarragain@ucc.ie

[3] The University of Manchester, Manchester, UK, {carole.goble,soiland-reyes}@manchester.ac.uk

This presentation will introduce a new specification for describing and distributing research datasets as re-usable, objects: **RO-Crate**. The work is an amalgam of the DataCrate specification and Research Object and is intended to distill a number of community efforts into a single easy-to implement specification for describing datasets at rest, on the web, and
packaged for distribution using standard mechanisms such as BagIt.

DataCrate [10] is a specification for packaging research data with both human and machine readable metadata, and has matured to a version 1.0 release . **Research Objects** (RO) provide 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 transcend 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 [2] as key components are often not sufficiently tracked, archived or reported.

BACKGROUND

Multiple data packaging initiatives have recently emerged, within the Research Data Alliance, Force11, DataOne and elsewhere; for example 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 any kind of dataset. Common among these is the use of 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.

Research Objects [1] are built on Linked Data standards: W3C RDF, JSON-LD, OAI-ORE, W3C Web Annotations, PROV, Dublin Core Terms, ORCID as well as the RO ontologies [3]. The RO Hub portal [4] uses RDF REST resources; and Research objects can be bundled as ZIP files [5] or Big Data BagIt archives [6, 7].

DATA PACKAGING PRINCIPLES

An RDA meeting on data packaging concluded that many initiatives have converged on similar solutions: simple folder structure; JSON-LD manifest; use of the schema.org vocabulary for core metadata; BagIt for fixity; OAI-ORE for aggregation but there is no single widely accepted standard for using these technologies. RO-Crate is proposed as a format for data packaging and dataset description designed to easy to add to repositories and archives as well as active research tools, compatible with Google's Dataset search, and that still allows communities to build domain-specific solutions. Frictionless data [8] could arguably fill this gap, with mature specifications, however as a simple JSON format with no formal extension mechanism it does not fully apply Linked Data principles, and would be harder to use in FAIR [13] integrations and extensions.

The new specification RO-Crate, will be based around these principles: a) all metadata is 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 and other packaging tools and compatible with digital preservation approaches; h) hooks to reuse existing domain formats; i) designed for easy 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-Crate is a fresh initiative, bringing together data archive and repository maintainers with existing Research Object, workflow and provenance communities. The RO-Crate working group started as a small core drawn mainly from the Research Object and RDA community. 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.

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.

REFERENCES

[1] 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

[2] Monya Baker (2016): 1,500 scientists lift the lid on reproducibility. *Nature* 5 33. https://doi.org/10.1038/533452a

[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):e0213013 https://doi.org/10.1371/journal.pone.0213013

[7] Farah Zaib Khan et al (2019): **Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv**. Submitted to *GigaScience*. https://doi.org/10.5281/zenodo.3196309

[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 [ACM Queue postprint available]

[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 2019-05-02]

[13] M. D. Wilkinson *et al.*,(2016) : **The FAIR Guiding Principles for scientific data management and stewardship,** *Scientific Data*, vol. 3, p. 160018, Mar. 2016. http://dx.doi.org/10.1038/sdata.2016.18