Research Object Community Update

Carole Goble, Stian Soiland-Reyes, Sean Bechhofer

School of Computer Science, The University of Manchester, Manchester, M13 9PL, UK
{carole.goble,soiland-reyes,sean.bechhofer}@manchester.ac.uk

Abstract

We highlight recent developments and approaches in the ResearchObject.org community. Research Object, originally proposed in [1] and since developed and expanded on [2][3], is a framework by which the many, nested and contributed components of research can be packaged together in a systematic way, and their context, provenance and relationships richly described. Research Objects (ROs) define ontology-based mechanisms for richly describing the contents of container manifests, and the relationships between them.

Update on the specification

In brief, a Research Object has two major components:

- A **container mechanism** for the components (embedded or externally referenced). Containers generally take two forms: (i) general portable packaging frameworks such as Docker, BagIt, Zip, Singularity and Conda; and (ii) bespoke platforms whereby the RO model is incorporated in the infrastructure, such as the ROHub [4]. Container profiles map the manifest metamodel to the container implementation, such as for BagIt [5] and for zip [6].
- A **manifest of metadata** that identifies and aggregates content (local and external) and support the attribution and provenance of each resource for credit and right versions. Rich annotations facilitate the use, re-use and interpretation of its contents, and provides an extensibility point for community-driven standards. Manifests are constructed by using established web standards, infrastructure and identifier schemes expressed as Linked Data or JSON. A collection of vocabularies and ontologies define the metamodel [2] used to describe the aggregation structure of the ROs. Manifest content profiles use domain ontologies to provide the application, type or task-specific content of those ROs: notably checklists of what should be there but also the provenance of where it came from; versions tracking evolution; and dependencies of what else is needed.

In this presentation we outline features of the latest specifications (https://w3id.org/ro/2016-01-28) and highlight approaches to container profiles for BagIt [5]. We highlight developing work on general approaches to encoding manifest content profiles using Linked Data and RDF Shapes in order to support general validation tools (https://github.com/researchobject/ro-show). Finally we will introduce alignments with key metadata discovery mark-up schemas BioSchemas (http://bioschemas.org/) and DCAT2 (https://www.w3.org/TR/vocab-dcat-2).

Update on the Community adoption

Research Objects are designed to: be tailored to be domain or type specific; work at many levels of granularity, with their own identifiers, citation metadata; and be snapshots or living to suit their place in the research lifecycle. The potential benefits of rich metadata coupled with portable containers include:

- **Exchange & Commons**: The transfer of knowledge, data and results between the services and actors and the development of RO Commons to enable reuse and sharing. References to remote content allows for access management due to privacy restrictions or data scales.
- **Preservation**: Snapshots of state of a collection of resources for fixed-point publishing.
- **Reproducibility**: Describing the structured collection, its components and its context in a rich enough way to support inspection and interpretation by people, and re-execution and comparison by computational machinery.
- **Active "release" oriented research**: Accumulating metadata to reflect versions, new configurations of content, evolutions, relationships between objects, and metadata reflecting who has added to the object or used it.

	Sponsors	**Discipline**	**Type**	**Exchange & Commons**	**Preservation**	**Reproducibility & Execution**	**Active "release"**
Workflows [3]. Common Workflow Language [7 ,8]	ELIXIR EOSCPilot BioExcel CWL	Any *Piloted in Life Science, Biomedicine & Astronomy*	Workflows	☑	☑	☑	☑
Big Data management. BDBags [5, 9]	NIH Data Commons	Any *Piloted in Biomedicine*	Big Data	☑			
STELAR Asthma eLab [10]	Farr Institute UK	Biomedicine	Data & Stats Analysis	☑	☑	☑	
GigaDB [11] SOAPDenovo2 pipelines [12]	GigaScience	Publishing Life Sciences	Data & Workflows	☑	☑		
SysBio packaging [13]	FAIRDOM	Systems Biology	Data & Models	☑			☑
BioCompute Objects [14]	FDA	Biomedicine	Workflows	☑		☑	
Metagenomic pipelines [15]	ELIXIR	Life Science	Workflows	☑		☑	
Earth Science. EVER-EST [16]	EPOS	Earth Science	Data & Workflows	☑	☑	☑	☑

Table 1: Snapshot of RO community efforts

Since 2010 a RO community has emerged to build on the specifications, for different disciplines, types and tasks. A summary of some of the work is given in Table 1. ROs were developed in full for workflow preservation in the EU project Wf4ever (http://www.wf4ever.org). That driver still dominates, as do the Life Sciences and Biomedical disciplines.

In the presentation we highlight trends in the community and identify gaps, notably in general tooling for RO construction, validation and viewing. We also preview RO-based standardisation efforts such as the IEEE P2791 BioCompute Working Group (http://sites.ieee.org/sagroups-2791/).

References

[1] Bechhofer S, De Roure D, Gamble M, Goble C and Buchan I (2010): **Research Objects: Towards Exchange and Reuse of Digital Knowledge.** *The Future of the Web for Collaborative Science* (FWCS 2010), Raleigh, USA, 1 April 2010. https://doi.org/10.1038/npre.2010.4626.1

[2] Bechhofer S, Buchan I, De Roure D, Missier P, Ainsworth J, Bhagat J, Couch P, Cruickshank D, Delderfield M, Dunlop I, Gamble M, Michaelides D, Owen S, Newman D, Sufi S, Goble C (2013): **Why linked data is not enough for scientists.** *Future Generation Computer Systems* **29**(2):599-611. https://doi.org/10.1016/j.future.2011.08.004

[3] Belhajjame K, Zhao J, Garijo D, Gamble M, Hettne K, Palma R, Mina E, Corcho O, Gómez-Pérez JM, Bechhofer S, Klyne G, Goble C (2015): **Using a suite of ontologies for preserving workflow-centric research objects.** *Web Semantics: Science, Services and Agents on the World Wide Web* **32**: 16-42 https://doi.org/10.1016/j.websem.2015.01.003

[4] Gómez-Pérez JM, Palma R, Garcia-Silva A (2017): **A Towards a Human-Machine Scientific Partnership Based on Semantically Rich Research Objects**. *Proc IEEE 13th International Conference on e-Science*, 24-27 Oct. 2017, Auckland, New Zealand https://doi.org/10.1109/eScience.2017.40

[5] Chard K, D'Arcy M, Heavner B, Foster I, Kesselman C, Madduri R, Rodriguez A, Soiland-Reyes S, Goble C, Clark K, Deutsch EW, Dinov I, Price N, Toga A (2016): **I'll Take That to Go: Big Data Bags and Minimal Identifiers for Exchange of Large, Complex Datasets**. *IEEE International Conference on Big Data 2016*. https://doi.org/10.1109/BigData.2016.7840618

[6] Soiland-Reyes S, Gamble M, Haines R (2014): **Research Object Bundle 1.0**. researchobject.org specification. *Zenodo*. https://w3id.org/bundle/2014-11-05/ http://doi.org/10.5281/zenodo.12586

[7] Khan FZ, Soiland-Reyes S, Sinnott RO, Lonie A, Crusoe MR (2018): **CWLProv – Interoperable retrospective provenance capture and its challenges**. *19th Bioinformatics Open Source Conference (BOSC 2018)*, Portland Oregon, USA. https://doi.org/10.7490/f1000research.1115721.1

[8] Robinson M, Soiland-Reyes S, Crusoe MR, Goble C (2017): **CWL Viewer: The Common Workflow language viewer.** *18th Annual Bioinformatics Open Source Conference (BOSC 2017)*, Prague, Czech Republic. https://doi.org/10.7490/f1000research.1114375.1

[9] Madduri RK, Chard K, D'Arcy M, Jung SC, Rodriguez A, Sulakhe D, Deutsch EW, Funk C, Heavner B, Richards M, Shannon P, Glusman G, Price N, Kesselman C, Foster I (2018): **Reproducible big data science: A case study in continuous FAIRness.** *bioRxiv* 268755. https://doi.org/10.1101/268755

[10] Custovic A, Ainsworth J, Arshad H, Bishop C, Buchan I, Cullinan P, Devereux G, Henderson J, Holloway J, Roberts G, Turner S, Woodcock A, Simpson A (2015): **The Study Team for Early Life Asthma Research (STELAR) consortium 'Asthma e-lab': team science bringing data, methods and investigators together.** *Thorax* **70**:799-801. https://doi.org/10.1136/thoraxjnl-2015-206781

[11] Edmunds, SC, Li, P, Hunter, CI, Zhe XS, Davidson RL, Nogoy N, Goodman L (2017): **Experiences in integrated data and research object publishing using GigaDB.** *International Journal on Digital Libraries* **18**: 99. https://doi.org/10.1007/s00799-016-0174-6

[12] González-Beltrán A, Li P, Zhao J, Avila-Garcia MS, Roos M, Thompson M, van der Horst E, Kaliyaperumal R, Luo R, Lee, TL, Lam TW, Edmunds SC, Sansone SA, Rocca-Serra P (2015) **From Peer-Reviewed to Peer-Reproduced in Scholarly Publishing: The Complementary Roles of Data Models and Workflows in Bioinformatics.** *PLoS ONE* **10**(7): e0127612. https://doi.org/10.1371/journal.pone.0127612

[13] Wolstencroft K, Krebs O, Snoep JL, Stanford NJ, Bacall F, Golebiewski M, Kuzyakiv R, Nguyen Q, Owen S, Soiland-Reyes S, Straszewski S, van Niekerk DD, Williams, AR, Malmström L, Rinn B, Müller W, Goble C (2017): **FAIRDOMHub: a repository and collaboration environment for sharing systems biology research.** *Nucleic Acids Research* **45**(D1):D404–D407. https://doi.org/10.1093/nar/gkw1032

[14] Alterovitz G, Dean II DA, Goble C, Crusoe MR, Soiland-Reyes S et al (2017): **Enabling Precision Medicine via standard communication of NGS provenance, analysis, and results**. *bioRxiv*. https://doi.org/10.1101/191783

[15] Meyer F, Finn R, Gerlach W, Mitchell A, Harrison T, Wilke A (2018). **On the way to research objects for environmental genomics (or metagenomics)**. *Zenodo* preprint, submitted to RO2018. http://doi.org/10.5281/zenodo.1309962

[16] Garcia-Silva A, Palma R, Gomez-Perez JM (2017): **Ensuring the Quality of Research Objects in the Earth Science Domain.** *Proc IEEE 13th International Conference on e-Science*, 24-27 Oct. 2017, Auckland, New Zealand. https://doi.org/10.1109/eScience.2017.62

Acknowledgements

We are grateful for funding from the **European Commission** for the **Horizon 2020** (H2020) projects BioExcel CoE (H2020-EINFRA-2015-1 675728), EOSCPilot (H2020-INFRADEV-2016-2 739563), ELIXIR-EXCELERATE (H2020-INFRADEV-1-2015-1 676559), and the **Seventh Framework** (FP7) project Wf4Ever (FP7-ICT-2009-6 270192). We would like to thank the Common Workflow Language project and the ResearchObject.org community.