Research Object Composer: A Tool for Publishing Complex Data Objects in the Cloud

Anita de Waard, Marina Soares E Silva,
Elsevier, Amsterdam

Abstract:

In recent years there has been a surge of efforts in bioinformatics to move data towards a 'Data Commons', (see e.g. [1 – 3]) that intends to "collate data with cloud computing infrastructure and commonly used software services, tools, and applications to create biomedical resources for the large-scale management, analysis, harmonization, and sharing of biomedical data" [3]. To create such a shared knowledge infrastructure, it is necessary that data and software, previously installed or hosted locally, become globally accessible, and made Findable, Accessible, Interoperable and Reusable [4]. As a first step to finding data and software, a unique, global, persistent identifier system is required, which is flexible enough to accommodate the multifarious inputs and outputs which a Data Commons can produce.

In earlier work, we developed an open source tool that allows users to mint a dataset DOI in an open data repository (Mendeley Data, http://data.mendeley.com), through an API that can be accessed from a proprietary cloud-based bioinformatics workflow tool [5]. However, this system did not allow for the minting of the complex datasets such as those generated on interconnected systems, which can include novel and existing multi-modal components, owned or created by third parties or by the user, and containing linear connections that allow insight into entire workflows. The Research Object model pioneered by Goble et al offers an ideal format to represent such complex objects [6]. Research Objects offer the ability to connect collections of data, code and workflows and describe them with a unique metadata 'Manifest', which allow the user insight in the entire workflow of the underlying set of knowledge artefacts [7].

We have developed a Research Object Composer (ROC) tool, that is able to generate a Research Object from a workflow and set of inputs. In our talk, we will describe the motivation of this work, the overall architecture of the system, and provide a brief demo of the current iteration of the Research Object Composer and offer an outlook on the future of this effort and its applicability to other scientific domains.

References:

[1] Grossman RL, Heath A, Murphy M, Patterson M, Wells W, A Case for Data Commons: Toward Data Science as a Service, Comput Sci Eng. 2016 Sep-Oct;18(5):10-20. doi: 10.1109/MCSE.2016.92. Epub 2016 Aug 24.

[2] Vivien R. Bonazzi, Philip E. Bourne, Should biomedical research be like Airbnb? PLoS Bioinformatics, Published: April 7, 2017 https://doi.org/10.1371/journal.pbio.2001818

[3] Robert L. Grossman, Data Lakes, Clouds, and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data, Trends in Genetics, Volume 35, Issue 3, March 2019, Pages 223-234, https://doi.org/10.1016/j.tig.2018.12.006

[4] Mark D. Wilkinson et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data volume 3, Article number: 160018 (2016)

[5] http://smart-api.info/ui/bf9abe9c17c9c78c432832382ef9e16a#/

[6] Sean Bechhofer, Iain Buchan, David De Roure, Paolo Missier, John Ainsworth, Jiten Bhagat, Phillip Couch, Don Cruickshank, Mark Delderfield, Ian Dunlop, Matthew Gamble, Danius Michaelides, Stuart Owen, David Newman, Shoaib Sufi, Carole Goble (2013) Why Linked Data is Not Enough for Scientists, Future Generation

Computer Systems 29(2), February 2013, Pages 599-611, ISSN 0167-739X, https://doi.org/10.1016/j.future.2011.08.004

[7] Khalid Belhajjame, Jun Zhao, Daniel Garijo, Matthew Gamble, Kristina Hettne, Raul Palma, Eleni Mina, Oscar Corcho, José Manuel Gómez-Pérez, Sean Bechhofer, Graham Klyne, Carole Goble (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