Nextflow integration for the Research Object Specification

Edgar Garriga Nogales[1,2], Paolo Di Tommaso[1], Cedric Notredame[1]

[1] Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Dr. Aiguader 88, 08003 Barcelona, Spain

[2] Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain

Introduction

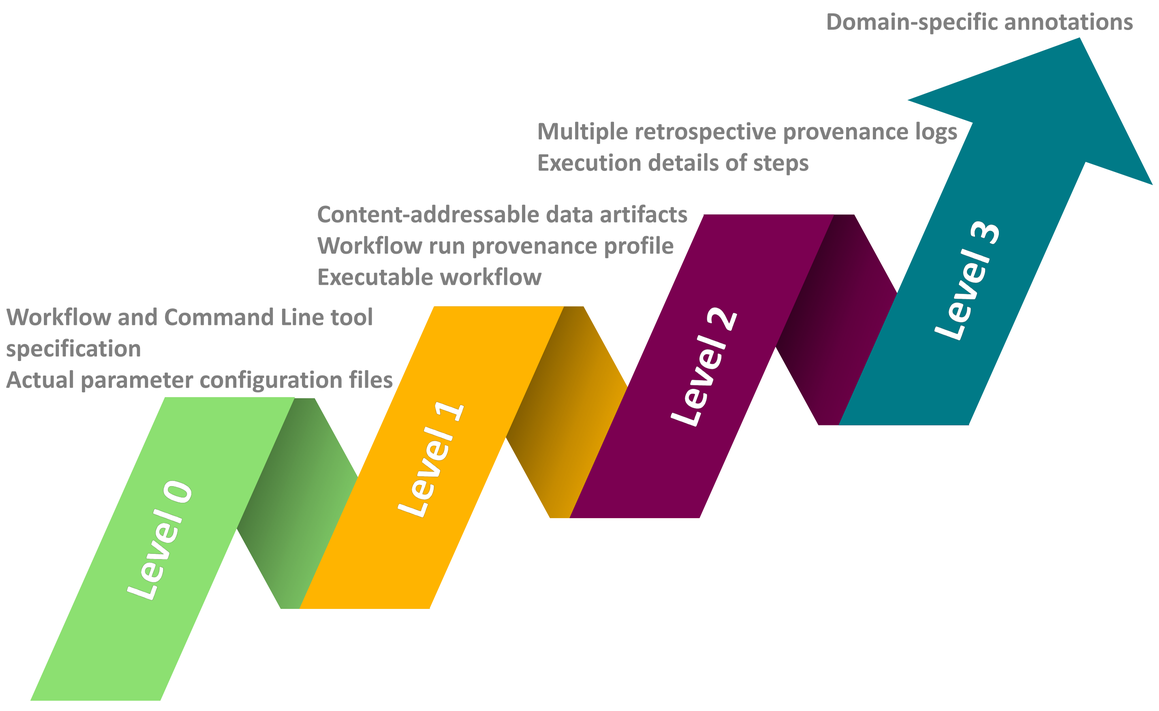
Nextflow is a framework based on the data flow programming model, which simplifies the writing of parallel and distributed pipelines while allowing developers to focus on the application. One of the strength of Nextflow is the integration of Docker and Singularity, allowing self-contained and **reproducible** computational pipelines. Another of the benefits of Nextflow is to be **polyglot**, and therefore able to deploy pipeline in any user chosen language such as Bash, C++, Python, etc.

Research Object is an approach bringing scientific computation one step closer to the FAIR concept. The main three objectives are: **Identity**, providing a unique identifier to the project, like the DOI. **Aggregation**, allowing the author to wrap all the useful elements for the project, slides, the article, etc.With the Research Object, we can share all the elements of a project together with the same ID. And finally, the last main principle is the **Annotation**, it means to provide an extra metadata to know the relation between elements, when and how they were produced.

Why Nextflow + ResearchObject

Once we achieved the **reproducibility** with Nextflow and containerization. Next challenge is sharing or distributing the workflow with the multiple files (workflow, scripts, notebooks, data, manuscript, etc.) as a **single referenceable entity**.
Or being able to understand the small details of the result generation like the exact command line/script in a the container, how long it took and who ran the pipeline are some of the issues for the personalized medicine of the near future.
Thanks to the combination of Nextflow and Research Object we are able to produce a single file with all the resources of the workflow. With the information of the **metadata** (author, doi, etc.), the **log** of the previous executions and the **provenance** of the intermediate files and the final results.

Provenance



Structure

The integration will make Nextflow able to produce a zip file with the following structure:
-A ***Data*** folder, with all the input files of our pipeline.
-The ***Workflow*** folder containing all the files of the pipeline base directory.
-The ***Snapshot*** folder is the one used re-execute the pipeline with the same parameters if its needed.
-The ***Metadata*** folder contains the log file of the past executions, the metadata file with information about the container used and the Nextflow version. The metadata folder contains the **provenance** file too.
-The ***Output*** folder, where the output of the execution is stored.
-The ***RO*** folder, containing RO's manifest with information about the author and the creation of the RO.

Results

The Nextflow-ResearchObject integration allows the creation of an ResearchObject package when executing a Nextflow workflow. The result is a **single package** with a unique identification (**identity**) containing all the important files of the project (**aggregation**) like the metadata, logs, results and the workflow directory.
Another important value of this integration is the generation of the **provenance** annotation. This information makes it easier to share and reproduce an analysis. It also increases the transparency on the procedure behind the analysis. These three elements contribute towards the three core principles of ResearchObject.