opendata.fit

Australian Characterisation Commons at Scale
imagingtools.org.au

A data analysis and publication platform for the web

Varvara Efremova, James Wilmot, Pall Thordarson
opendatafit@proton.me

UNSW SYDNEY

What is opendata.fit?

opendata.fit is a cloud-based web platform and execution engine for use by researchers to manage analysis workflows and datasets from development to publication, with a focus on model fitting and optimisation.

Motivation

Model optimisation

Existing analysis platforms (Galaxy, EcoCommons, etc) are domain-specific & largely focused on data processing workflows.
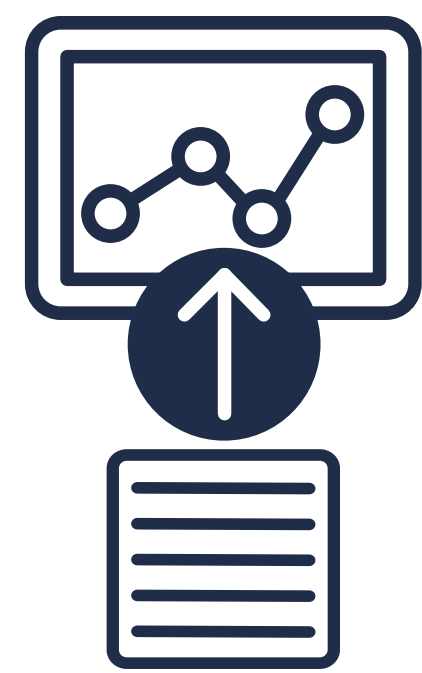
No existing platforms target model fitting and optimisation methods which are used in a multitude of fields (drug discovery, materials science, astronomy), forcing researchers to rely on bespoke solutions.

Reproducibility

Existing analysis workflows vary widely between groups and can be difficult to track and publish in their entirety, resulting in lack of reproducibility.

opendata.fit aims to improve reproduciblity by encapsulating entire analysis workflows including history into a single JSON file format. Complete, reproducible workflows can then be included in a publication by citing a single link.

Upload instrument data

Select from a library of analysis algorithms

Perform analysis using a simple web interface

Visualise results

Write and contribute custom analysis algorithms

Publish complete workflow with a single link

[42] Jane Smith, opendata.fit: Example Analysis, Sept 2022. Available at https://app.opendata.fit/s/cfdf85
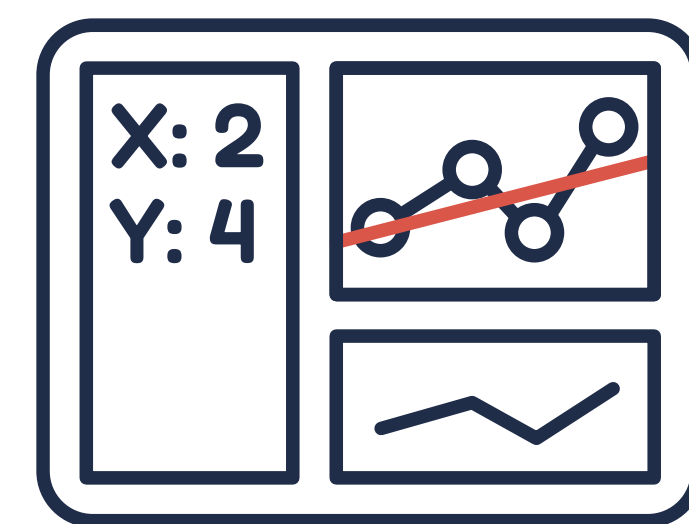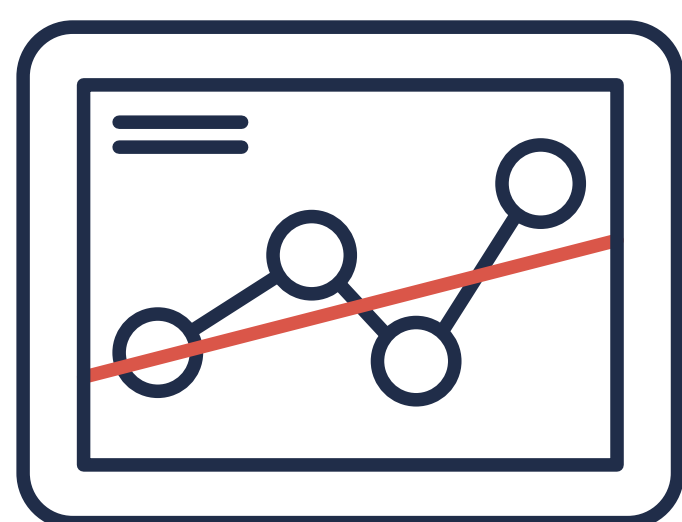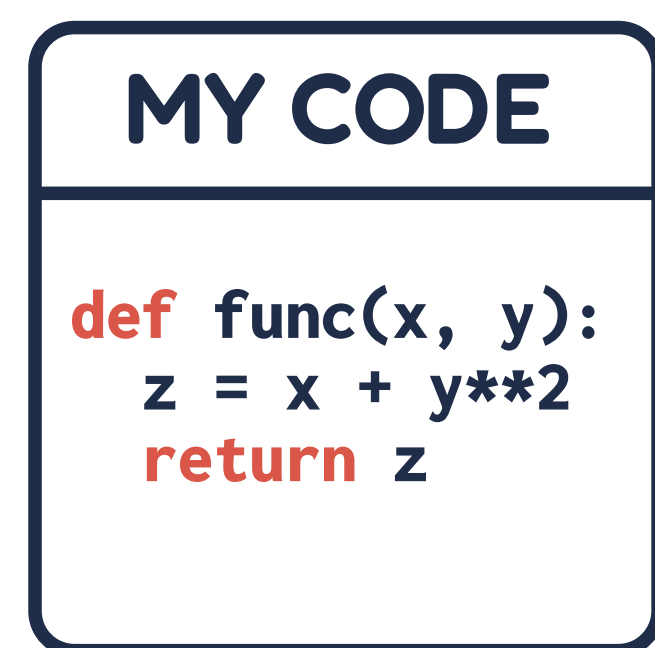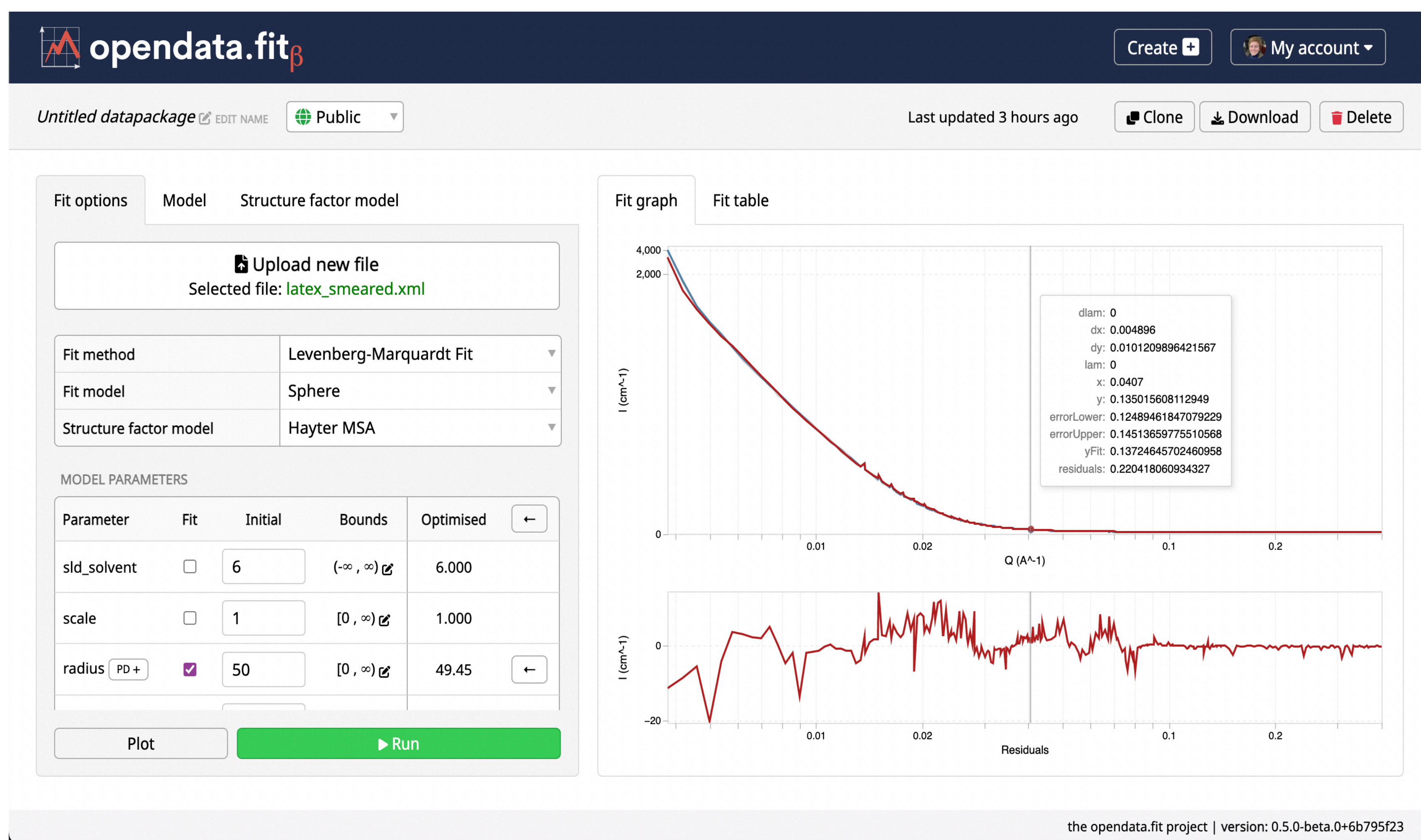
```
{
  "title": "Example datapackage",
  "description": "Example of a datapackage encapsulating a workflow",
  "algorithms": [
    {
      "name": "example_algorithm",
      "title": "Example algorithm",
      "code": "def func(x, y):\n  z = x + y**2\n  return {\n    \"z\": z,\n  }",
      "inputs": [
        {
          "name": "x",
          "title": "X input parameter",
          "description": "A parameter to the algorithm (can be input data, an initial
                          parameter value, options for the algorithm, or any other value)",
          "type": "resource",
          "resource": "x_data"
        },
        ...
      ],
      "outputs": [...]
    }
  ],
  "resources": [
    {
      "name": "x_data",
      "title": "X input parameter",
      "format": "number",
      "data": 42
    },
    ...
  ],
  "displays": [
    {
      "name": "example_workflow_display",
      "title": "Example workflow display",
      "description": "Defines a display for the example workflow. Describes the user interface
                      components - input fields, output fields, visualisations, etc - to be
                      displayed in the web application.",
      ...
    }
  ],
  "views": [
    {
      "name": "params_view",
      "title": "Parameter input view",
      "resources": [
        "x_data",
        "y_data"
      ],
      "specType": "opendatafit-params"
    },
    ...
  ]
}
```

Datapackages

Workflows in opendata.fit are packaged into a single JSON file format called a datapackage, which extends the existing open source Frictionless data standard. Visualisations are defined using the open source Vega visualisation language.

Each datapackage is a complete, portable record of a workflow inluding datasets, algorithms, analysis process and input and output parameters. Datapackage algorithms are fully containerised through Docker maintaining a consistent execution environment.

Built on FAIR data principles

Findable — Datapackages and resources are created with a unique and persistent URL

Accessible — Datapackages can be retrieved via API or downloaded via the web interface

Interoperable — Datapackages extend the well documented Frictionless data standard

Reusable — Datapackages contain provenance metadata and full history of the analysis allowing it to be reused