Tutorial to manually initiate and fill the ARC (i.e. without the tools being developed)

May, 2021

Hajira Jabeen & Dominik Brilhaus

Disclaimer

Please note:

1. The DataPLANT tools are **work in progress**.

2. The following is a **manual** route towards building an ARC

3. This is for test / develop purposes and **subject to change**. Do not reformat all your datasets.
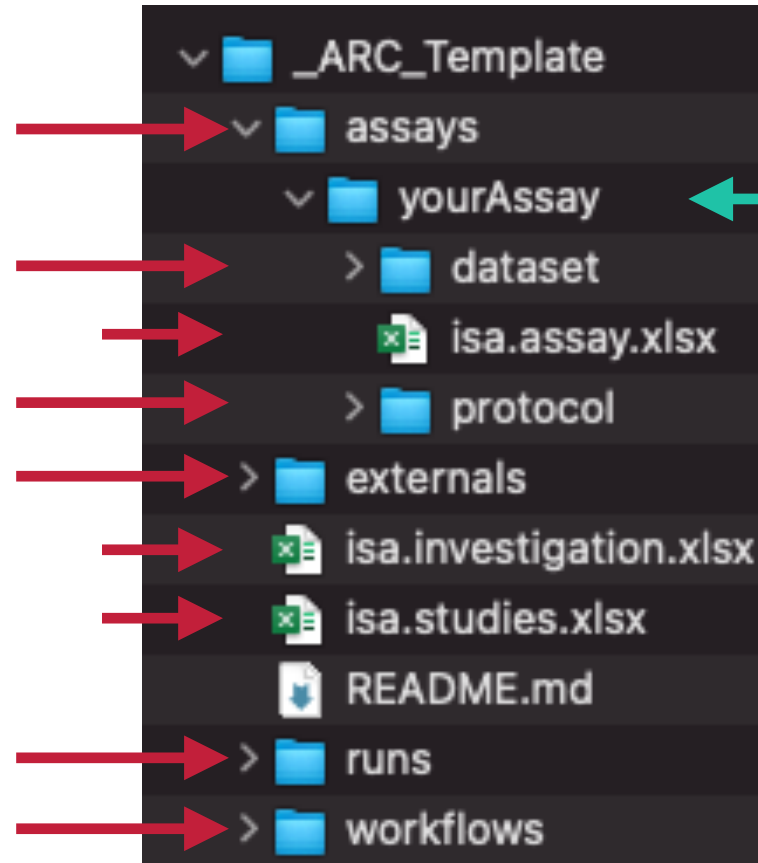
4. Talk to Hajira and Dominik first.

Please do not rename items with fixed names:

isa*.xslx files:
- isa.assay.xlsx
- isa.investigation.xlsx
- isa.studies.xlsx

ARC folders:
- assays/
- assays/*/dataset/
- assays/*/protocol/
- externals/
- runs/
- workflows/

```
∨ 📁 _ARC_Template
  ∨ 📁 assays
    ∨ 📁 yourAssay
      > 📁 dataset
        📊 isa.assay.xlsx
      > 📁 protocol
  > 📁 externals
  📊 isa.investigation.xlsx
  📊 isa.studies.xlsx
  📄 README.md
  > 📁 runs
  > 📁 workflows
```

Feel free to rename

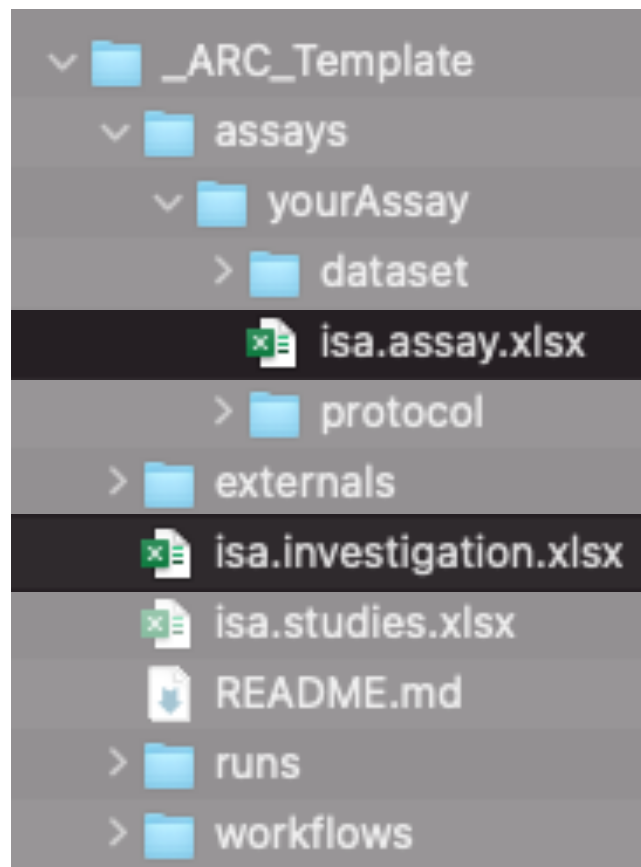Files and folders starting with your* (yourAssay, yourRun, etc.)

Feel free to add

Additional files and folders **anywhere** within the ARC. Just make sure not to give them one of the fixed names above.

Add project metadata

to "isa.investigation"

Describe your dataset(s)

in "isa.assay.xlsx"

Note: Ask your data stewards for a metadata template that is suitable with your assay
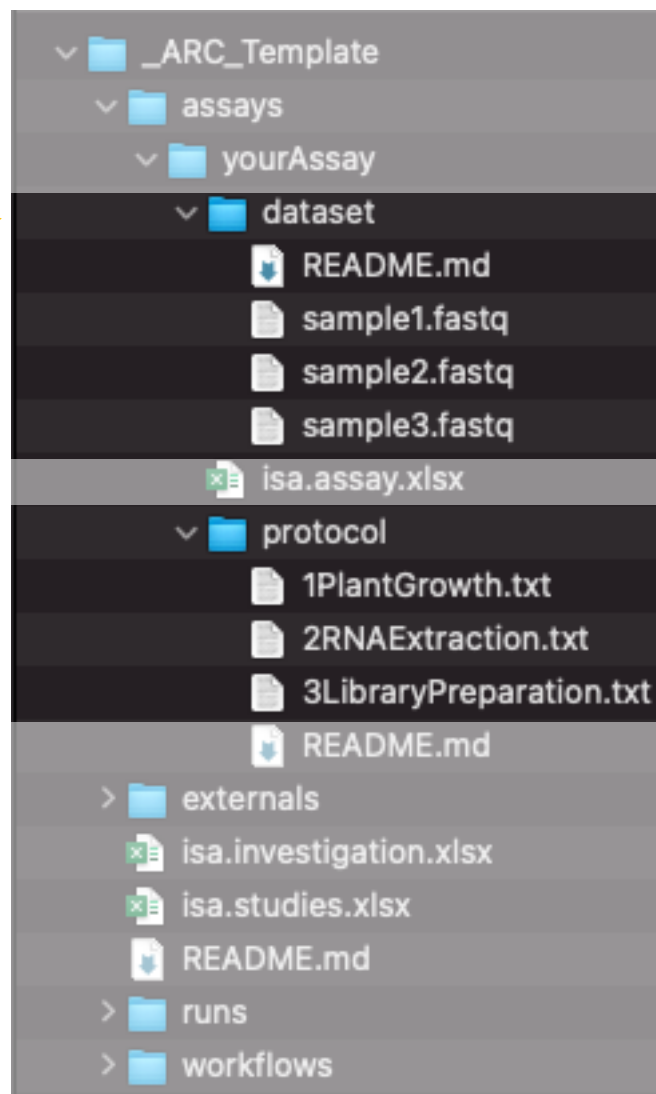
_ARC_Template
- assays
  - yourAssay
    - dataset
    - isa.assay.xlsx
    - protocol
- externals
- isa.investigation.xlsx
- isa.studies.xlsx
- README.md
- runs
- workflows

Add raw data files

to "dataset/"

Note: The names of the raw data files are listed in


```
∨ 📁 _ARC_Template
  ∨ 📁 assays
    ∨ 📁 yourAssay
      ∨ 📁 dataset
           📄 README.md
           📄 sample1.fastq
           📄 sample2.fastq
           📄 sample3.fastq
         📊 isa.assay.xlsx
      ∨ 📁 protocol
           📄 1PlantGrowth.txt
           📄 2RNAExtraction.txt
           📄 3LibraryPreparation.txt
         📄 README.md
    > 📁 externals
      📊 isa.investigation.xlsx
      📊 isa.studies.xlsx
      📄 README.md
  > 📁 runs
  > 📁 workflows
```

Add laboratory protocols

to "protocol/"

Note: Add all protocols required to reproduce the raw data files stored in "dataset/". These can be free-text protocols (e.g. coming from an electronic lab notebook) just as you would add them to your paper's "materials & methods" section.
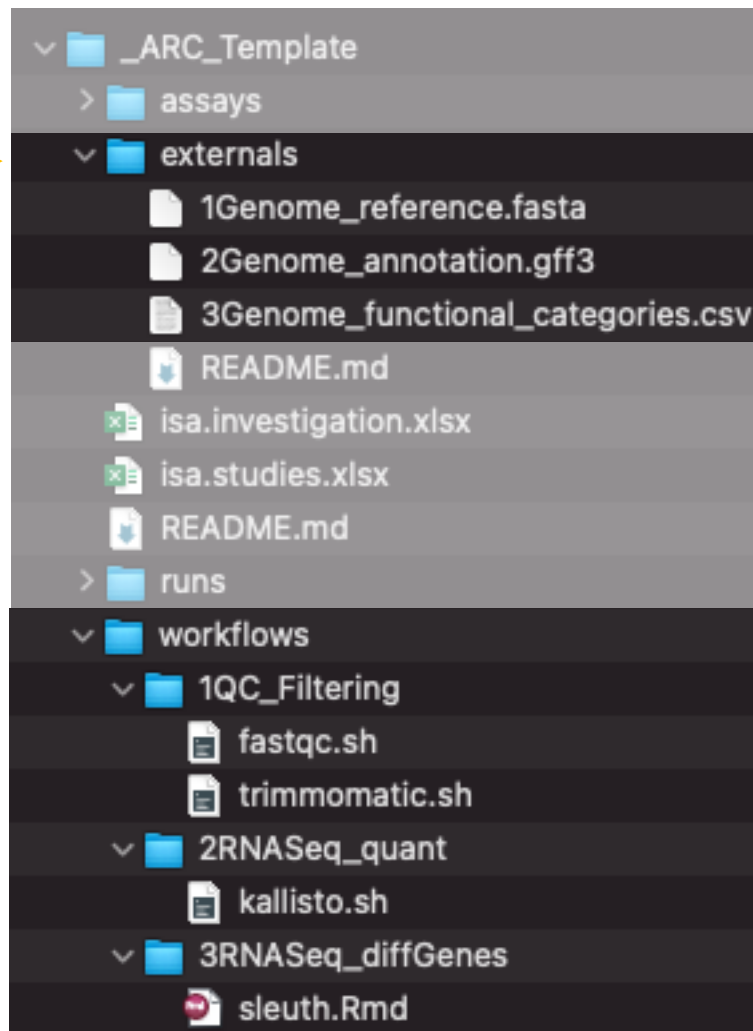
Add external data

to "externals/"

Note: This is additional data required for your analyses that typically originates from other projects / publications / databases.

Pro tipp: describe the origin, source and date or version of each file in the README.md

Add your analyses

to "workflows/"

Note: these are computations to process your data. Although typically scripts or code, feel free to add (descriptions of) alternative analyses.

- ∨ 📁 _ARC_Template
  - \> 📁 assays
  - ∨ 📁 externals
    - 📄 1Genome_reference.fasta
    - 📄 2Genome_annotation.gff3
    - 📄 3Genome_functional_categories.csv
    - 📄 README.md
  - 📊 isa.investigation.xlsx
  - 📊 isa.studies.xlsx
  - 📄 README.md
  - \> 📁 runs
  - ∨ 📁 workflows
    - ∨ 📁 1QC_Filtering
      - 📄 fastqc.sh
      - 📄 trimmomatic.sh
    - ∨ 📁 2RNASeq_quant
      - 📄 kallisto.sh
    - ∨ 📁 3RNASeq_diffGenes
      - 📄 sleuth.Rmd

… to be continued

Ressources

- CEPLAS
  - Homepage: https://www.ceplas.eu/en/research/data-science-and-data-management/
  - GitHub: https://github.com/CEPLAS-FAIRidise

- DataPLANT
  - Homepage: https://nfdi4plants.de