# Arcus Research
# Data Management

## RDM 101 + project template
## Arcus Labs
## Minimum requirements

template update: 2020 March 20
slides update: 2021 August 2

Children's Hospital of Philadelphia℠

# What is Research Data Management?

# Research Data Management (RDM) 101

- Research data: Information collected during the course of research processes used for analysis
- RDM: Process of organizing, annotating, preserving, and sometimes sharing research data in all its forms
- Applies to entire lifecycle of research data
- Requires ongoing iteration rather than one-time application
- Best practices/recommendations rather than prescriptions
  - Universal
  - Domain-specific

# Why do we care about RDM?

# Regulatory requirements

- Many funding agencies (CDC, EPA, DoD & more) require data sharing & pre-written data management plans
- NIH requires a data sharing plan for all grants over 500k
- NIH will require data sharing & management plan for **all** grants starting Jan 2023
- Many sub groups such as data consortiums or centers of interest require data sharing and management
- Many publishers require data sharing along with publication

# Benefits to the researcher

Well managed data:

- Is easier to quality check & validate during collection & analysis
- Reduces knowledge drain during staff turnover
- Reduces bottlenecks by shifting implicit knowledge to explicit knowledge
- Maximizes ease of reuse for future analysis & publications
- Reduces administrative burden for combining with additional datasets for expanded analysis
- Reduces administrative burden for sharing data & collaborating with new researchers
- Simplifies responding to questions or challenges arising from publication & review

# Benefits to the researcher

Research Data Management is about getting the most out of the hard work researchers are doing and the valuable data being collected

# Contextualizing research data

Understanding research as

- a process
- an environment
- a community

Documenting data as well as

- the objective of the research effort
- the methods of data collection
- the methods of data transformation
- the methods of data analysis

# Contextualizing research data

Research data

+

Access tools

+

Contextual files

=

Reproducibility

# Arcus Project Template

- File directory structure
- Organized but flexible
- Project-based instead of person-based
- Research data + access tools + contextual files + metadata
- Use for new, in-progress, on-going, and completed work
- Collaborative stewardship of data in Labs by Arcus and study teams

# Arcus Project Template

```
/home_dir/project_name/
├── README.md
├── configs
│   └── README.md
├── data
│   ├── README.md
│   ├── endpoints
│   ├── interim
│   ├── raw
│   └── ref-data
├── logs
├── manifests
│   ├── README.md
│   ├── file_derivation.csv
│   ├── file_manifest.csv
│   ├── participant-crosswalk.txt
│   ├── participant_family_role.csv
│   └── participant_manifest.csv
├── models
│   └── README.md
├── references
│   ├── README.md
│   └── protocols
├── reports
│   ├── README.md
│   ├── figures
│   ├── log.md
│   ├── methods.md
│   └── tables
├── requirements
│   ├── README.md
│   └── project_requirements.txt
└── src
    ├── README.md
    ├── notebooks
    │   └── README.md
    ├── rules
    ├── scripts
    └── tests
```

```
/home_dir/project_name/
├── README.md
├── configs
│   └── README.md
├── data
│   ├── README.md
│   ├── endpoints
│   ├── interim
│   ├── raw
│   └── ref-data
├── logs
├── manifests
│   ├── README.md
│   ├── file_derivation.csv
│   ├── file_manifest.csv
│   ├── participant-crosswalk.txt
│   ├── participant_family_role.csv
│   └── participant_manifest.csv
├── models
│   └── README.md
├── references
│   ├── README.md
│   └── protocols
├── reports
│   ├── README.md
│   ├── figures
│   ├── log.md
│   ├── methods.md
│   └── tables
├── requirements
│   ├── README.md
│   └── project_requirements.txt
└── src
    ├── README.md
    ├── notebooks
    │   └── README.md
    ├── rules
    ├── scripts
    └── tests
```

# Arcus Project Template

**Research data**

# Arcus Project Template

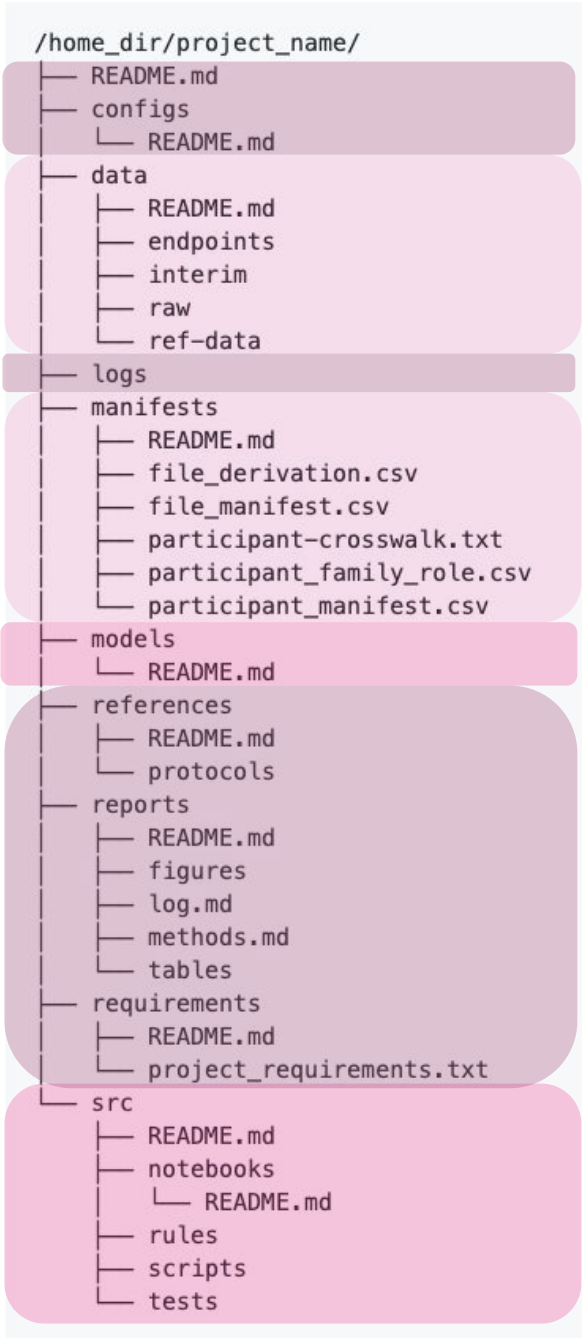
```
/home_dir/project_name/
├── README.md
├── configs
│   └── README.md
├── data
│   ├── README.md
│   ├── endpoints
│   ├── interim
│   ├── raw
│   └── ref-data
├── logs
├── manifests
│   ├── README.md
│   ├── file_derivation.csv
│   ├── file_manifest.csv
│   ├── participant-crosswalk.txt
│   ├── participant_family_role.csv
│   └── participant_manifest.csv
├── models
│   └── README.md
├── references
│   ├── README.md
│   └── protocols
├── reports
│   ├── README.md
│   ├── figures
│   ├── log.md
│   ├── methods.md
│   └── tables
├── requirements
│   ├── README.md
│   └── project_requirements.txt
└── src
    ├── README.md
    ├── notebooks
    │   └── README.md
    ├── rules
    ├── scripts
    └── tests
```

**Access tools**

# Arcus Project Template

```
/home_dir/project_name/
├── README.md
├── configs
│   └── README.md
├── data
│   ├── README.md
│   ├── endpoints
│   ├── interim
│   ├── raw
│   └── ref-data
├── logs
├── manifests
│   ├── README.md
│   ├── file_derivation.csv
│   ├── file_manifest.csv
│   ├── participant-crosswalk.txt
│   ├── participant_family_role.csv
│   └── participant_manifest.csv
├── models
│   └── README.md
├── references
│   ├── README.md
│   └── protocols
├── reports
│   ├── README.md
│   ├── figures
│   ├── log.md
│   ├── methods.md
│   └── tables
├── requirements
│   ├── README.md
│   └── project_requirements.txt
└── src
    ├── README.md
    ├── notebooks
    │   └── README.md
    ├── rules
    ├── scripts
    └── tests
```

**Contextual files**




# Arcus Project Template

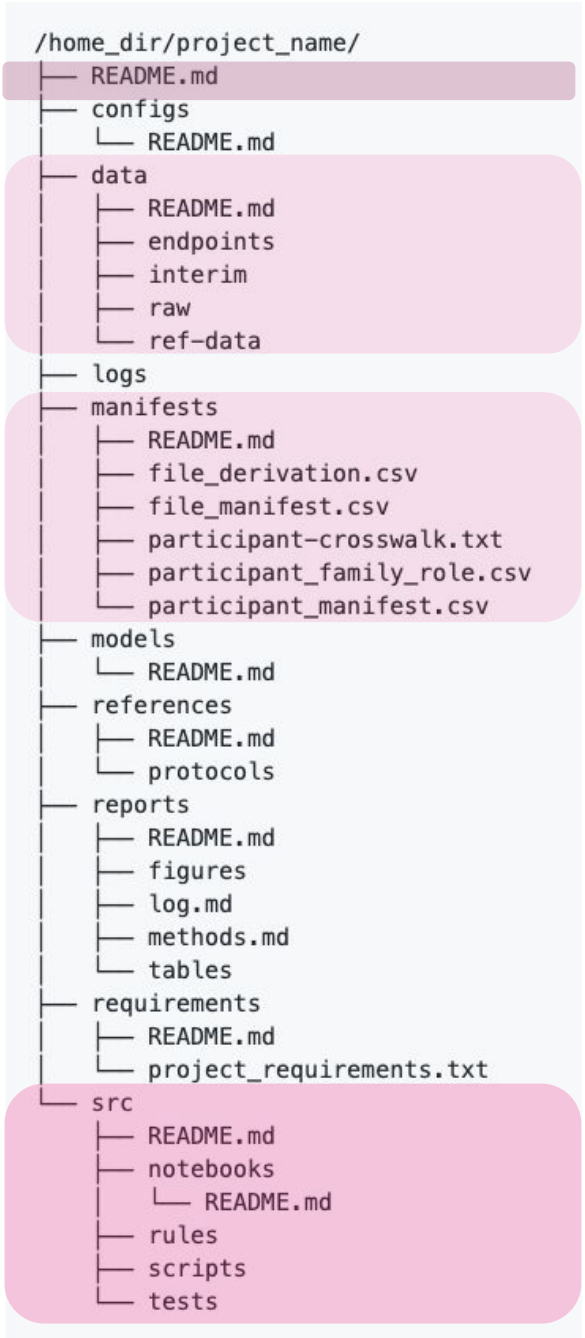
**Research data**

\+

**Access tools**

\+

**Contextual files**




# Arcus Project Template

**Research data**

\+

**Access tools**

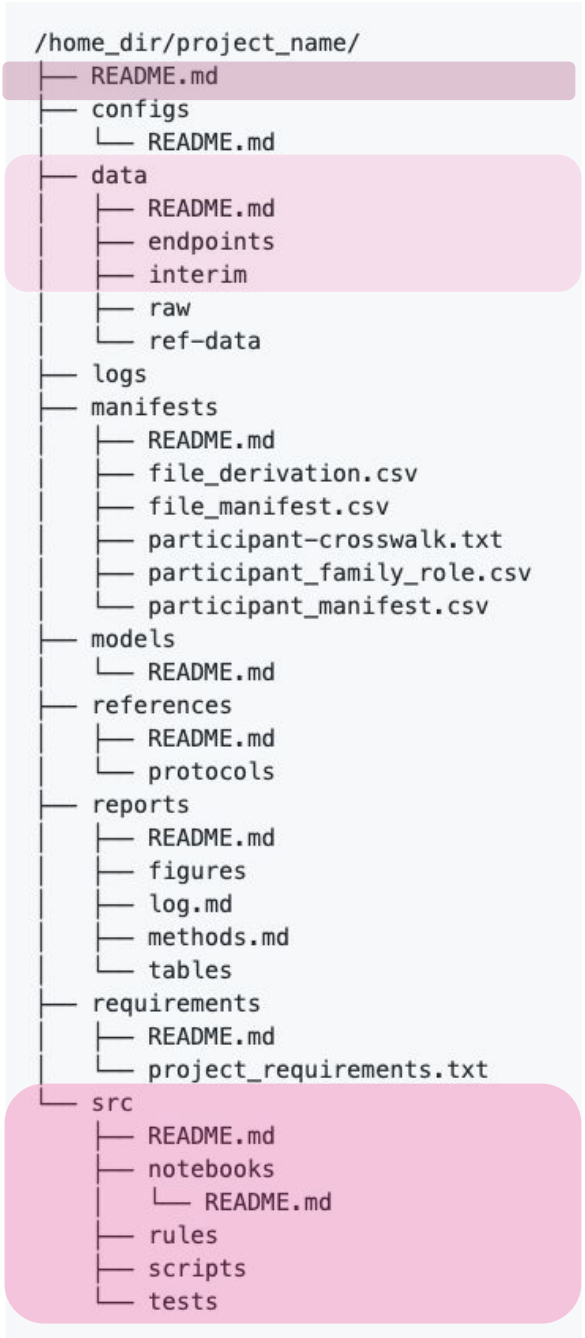
\+

**Contextual files**

**Minimum required components managed by Arcus & study team**



Children's Hospital of Philadelphia®




# Arcus Project Template

**Research data**

\+

**Access tools**

\+

**Contextual files**

**Minimum required components managed by study team**

```
/home_dir/project_name/
├── README.md
├── configs
│   └── README.md
├── data
│   ├── README.md
│   ├── endpoints
│   ├── interim
│   ├── raw
│   └── ref-data
├── logs
├── manifests
│   ├── README.md
│   ├── file_derivation.csv
│   ├── file_manifest.csv
│   ├── participant-crosswalk.txt
│   ├── participant_family_role.csv
│   └── participant_manifest.csv
├── models
│   └── README.md
├── references
│   ├── README.md
│   └── protocols
├── reports
│   ├── README.md
│   ├── figures
│   ├── log.md
│   ├── methods.md
│   └── tables
├── requirements
│   ├── README.md
│   └── project_requirements.txt
└── src
    ├── README.md
    ├── notebooks
    │   └── README.md
    ├── rules
    ├── scripts
    └── tests
```

# Arcus Project Template

**Research data**

\+

**Access tools**

\+

**Contextual files**

**Minimum required components managed by Arcus**

```
/home_dir/project_name/
├── README.md
├── configs
│   └── README.md
├── data
│   ├── README.md
│   ├── endpoints
│   ├── interim
│   ├── raw
│   └── ref-data
├── logs
├── manifests
│   ├── README.md
│   ├── file_derivation.csv
│   ├── file_manifest.csv
│   ├── participant-crosswalk.txt
│   ├── participant_family_role.csv
│   └── participant_manifest.csv
├── models
│   └── README.md
├── references
│   ├── README.md
│   └── protocols
├── reports
│   ├── README.md
│   ├── figures
│   ├── log.md
│   ├── methods.md
│   └── tables
├── requirements
│   ├── README.md
│   └── project_requirements.txt
└── src
    ├── README.md
    ├── notebooks
    │   └── README.md
    ├── rules
    ├── scripts
    └── tests
```

# Arcus Project Template

## Required components:

- parent folders
  - data/
  - manifests/ *
  - src/
- README
  - for overall project
  - all folders used

* maintained by Arcus

```
/home_dir/project_name/
├── README.md
├── configs
│   └── README.md
├── data
│   ├── README.md
│   ├── endpoints
│   ├── interim
│   ├── raw
│   └── ref-data
├── logs
├── manifests
│   ├── README.md
│   ├── file_derivation.csv
│   ├── file_manifest.csv
│   ├── participant-crosswalk.txt
│   ├── participant_family_role.csv
│   └── participant_manifest.csv
├── models
│   └── README.md
├── references
│   ├── README.md
│   └── protocols
├── reports
│   ├── README.md
│   ├── figures
│   ├── log.md
│   ├── methods.md
│   └── tables
├── requirements
│   ├── README.md
│   └── project_requirements.txt
└── src
    ├── README.md
    ├── notebooks
    │   └── README.md
    ├── rules
    ├── scripts
    └── tests
```

# Arcus Project Template

## Required practices:

- version control
- standardized naming conventions*
  - file names
  - folder names
  - project name
- retention
  - regular review of contents and files for long-term use/preservation and for deletion

*RDM resources available

```
/home_dir/project_name/
├── README.md
├── configs
│   └── README.md
├── data
│   ├── README.md
│   ├── endpoints
│   ├── interim
│   ├── raw
│   └── ref-data
├── logs
├── manifests
│   ├── README.md
│   ├── file_derivation.csv
│   ├── file_manifest.csv
│   ├── participant-crosswalk.txt
│   ├── participant_family_role.csv
│   └── participant_manifest.csv
├── models
│   └── README.md
├── references
│   ├── README.md
│   └── protocols
├── reports
│   ├── README.md
│   ├── figures
│   ├── log.md
│   ├── methods.md
│   └── tables
├── requirements
│   ├── README.md
│   └── project_requirements.txt
└── src
    ├── README.md
    ├── notebooks
    │   └── README.md
    ├── rules
    ├── scripts
    └── tests
```

# Arcus Project Template

**Contextual files**

```
/home_dir/project_name/
├── README.md
├── configs
│   └── README.md
├── data
│   ├── README.md
│   ├── endpoints
│   ├── interim
│   ├── raw
│   └── ref-data
├── logs
├── manifests
│   ├── README.md
│   ├── file_derivation.csv
│   ├── file_manifest.csv
│   ├── participant-crosswalk.txt
│   ├── participant_family_role.csv
│   └── participant_manifest.csv
├── models
│   └── README.md
├── references
│   ├── README.md
│   └── protocols
├── reports
│   ├── README.md
│   ├── figures
│   ├── log.md
│   ├── methods.md
│   └── tables
├── requirements
│   ├── README.md
│   └── project_requirements.txt
└── src
    ├── README.md
    ├── notebooks
    │   └── README.md
    ├── rules
    ├── scripts
    └── tests
```

# Arcus Project Template

## README *

- “homepage” document
- Context/guidance around files and processes in your research
- Document how teams manage, organize, use files
- Ease transitions between team members and over time

*RDM resource available

```
/home_dir/project_name/
├── README.md
├── configs
│   └── README.md
├── data
│   ├── README.md
│   ├── endpoints
│   ├── interim
│   ├── raw
│   └── ref-data
├── logs
├── manifests
│   ├── README.md
│   ├── file_derivation.csv
│   ├── file_manifest.csv
│   ├── participant-crosswalk.txt
│   ├── participant_family_role.csv
│   └── participant_manifest.csv
├── models
│   └── README.md
├── references
│   ├── README.md
│   └── protocols
├── reports
│   ├── README.md
│   ├── figures
│   ├── log.md
│   ├── methods.md
│   └── tables
├── requirements
│   ├── README.md
│   └── project_requirements.txt
└── src
    ├── README.md
    ├── notebooks
    │   └── README.md
    ├── rules
    ├── scripts
    └── tests
```

# Arcus Project Template

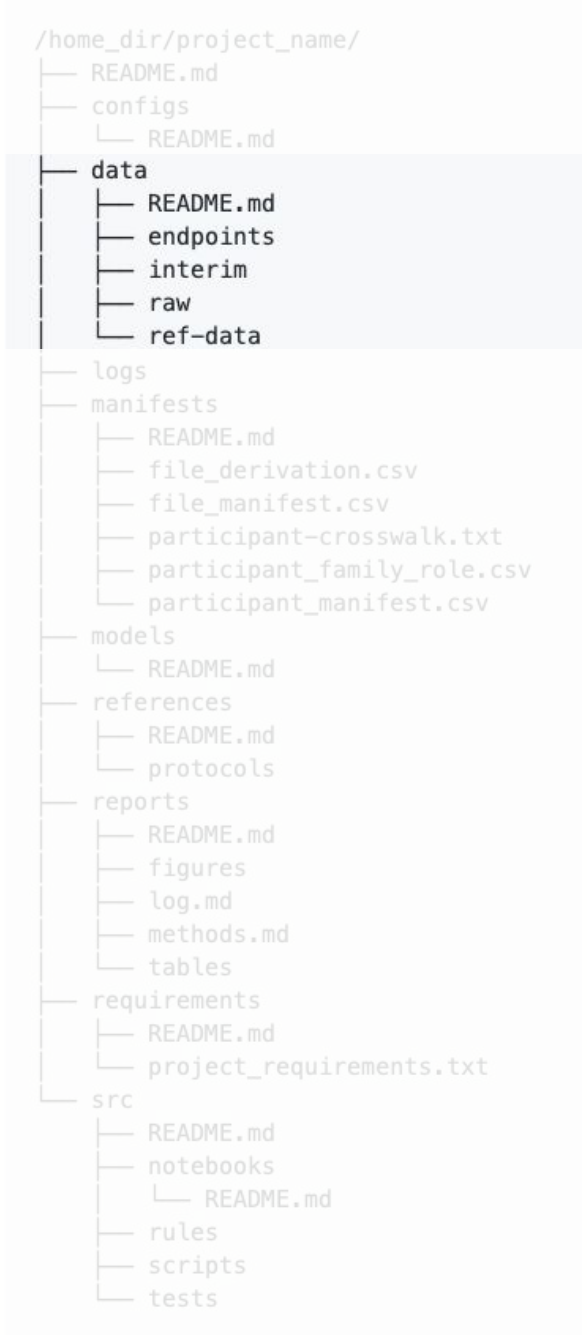
**README** contents

- people
- folder overview
- dates
- related files
- process information
- change log/version control
- naming conventions

```
/home_dir/project_name/
├── README.md
├── configs
│   └── README.md
├── data
│   ├── README.md
│   ├── endpoints
│   ├── interim
│   ├── raw
│   └── ref-data
├── logs
├── manifests
│   ├── README.md
│   ├── file_derivation.csv
│   ├── file_manifest.csv
│   ├── participant-crosswalk.txt
│   ├── participant_family_role.csv
│   └── participant_manifest.csv
├── models
│   └── README.md
├── references
│   ├── README.md
│   └── protocols
├── reports
│   ├── README.md
│   ├── figures
│   ├── log.md
│   ├── methods.md
│   └── tables
├── requirements
│   ├── README.md
│   └── project_requirements.txt
└── src
    ├── README.md
    ├── notebooks
    │   └── README.md
    ├── rules
    ├── scripts
    └── tests
```

# Arcus Project Template

**Research data**




# Arcus Project Template

## data/

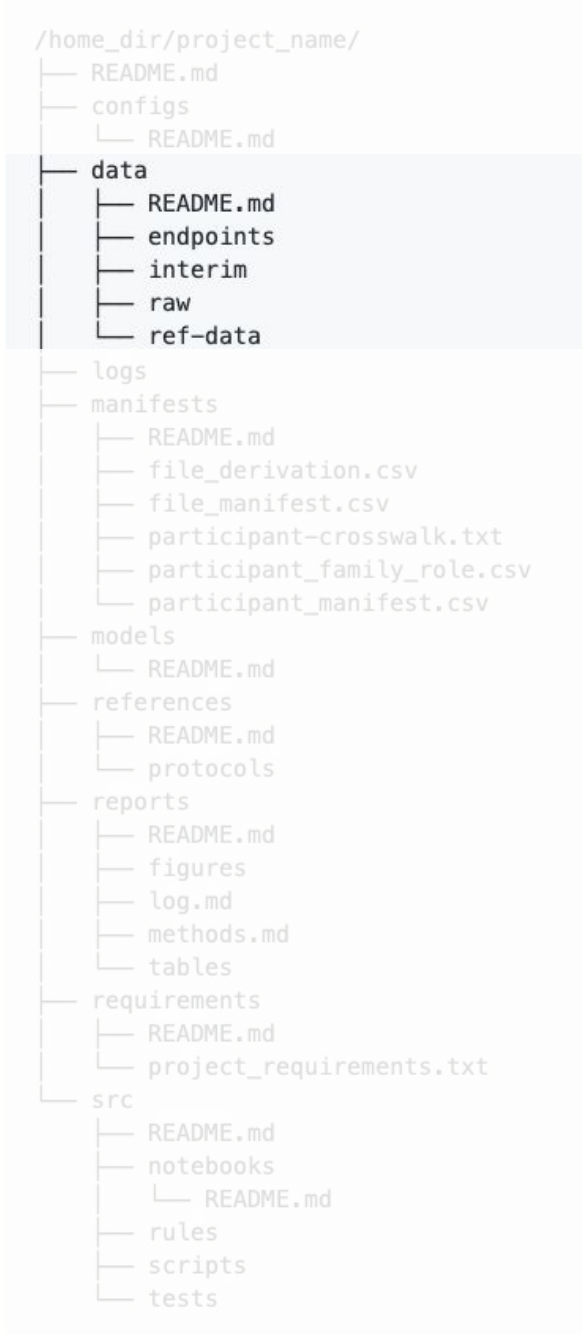
- Maintains descriptions of authoritative source data and their associated files and metadata in both raw and processed formats
- Required component
- Jointly managed by Arcus/study team

```
/home_dir/project_name/
├── README.md
├── configs
│   └── README.md
├── data
│   ├── README.md
│   ├── endpoints
│   ├── interim
│   ├── raw
│   └── ref-data
├── logs
├── manifests
│   ├── README.md
│   ├── file_derivation.csv
│   ├── file_manifest.csv
│   ├── participant-crosswalk.txt
│   ├── participant_family_role.csv
│   └── participant_manifest.csv
├── models
│   └── README.md
├── references
│   ├── README.md
│   └── protocols
├── reports
│   ├── README.md
│   ├── figures
│   ├── log.md
│   ├── methods.md
│   └── tables
├── requirements
│   ├── README.md
│   └── project_requirements.txt
└── src
    ├── README.md
    ├── notebooks
    │   └── README.md
    ├── rules
    ├── scripts
    └── tests
```

# Arcus Project Template

## data/raw/

- Arcus-delivered data goes here
- Study team generated data brought into Arcus goes here
- Authoritative source data that should never be deleted
- Organize in subdirectories if necessary




# Arcus Project Template

## data/interim

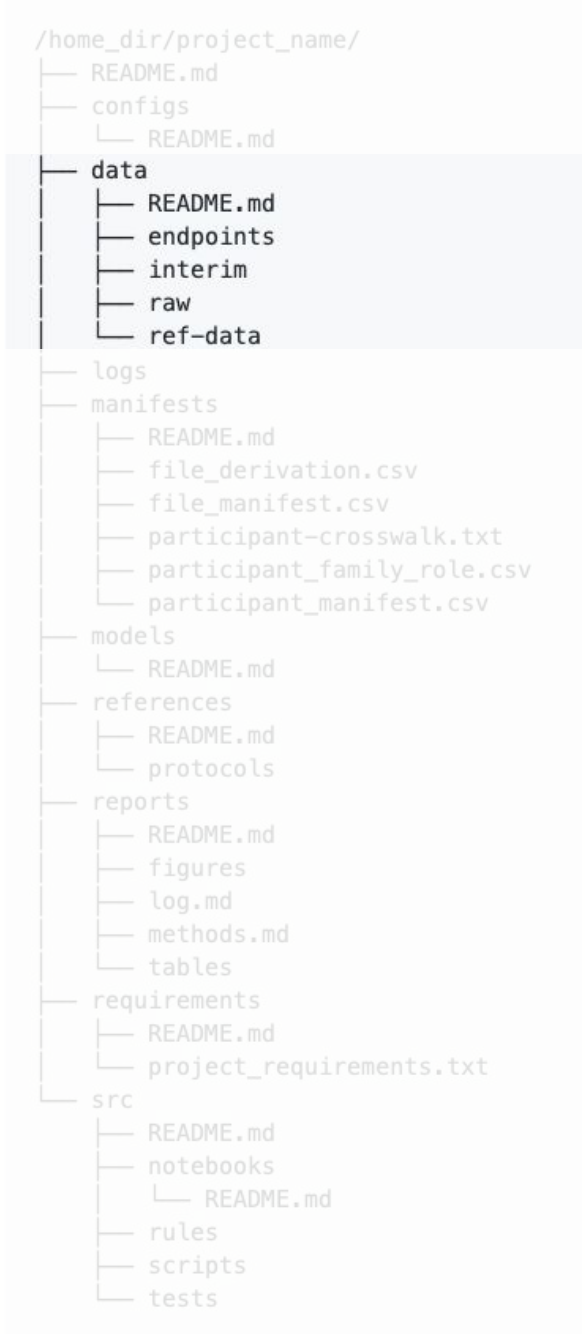
- Managed by study team
- Unregulated space for intermediate and temporary files
- Not necessary to save long-term
- Recommend establishing retention schedules for regular review/clean-up

```
/home_dir/project_name/
├── README.md
├── configs
│   └── README.md
├── data
│   ├── README.md
│   ├── endpoints
│   ├── interim
│   ├── raw
│   └── ref-data
├── logs
├── manifests
│   ├── README.md
│   ├── file_derivation.csv
│   ├── file_manifest.csv
│   ├── participant-crosswalk.txt
│   ├── participant_family_role.csv
│   └── participant_manifest.csv
├── models
│   └── README.md
├── references
│   ├── README.md
│   └── protocols
├── reports
│   ├── README.md
│   ├── figures
│   ├── log.md
│   ├── methods.md
│   └── tables
├── requirements
│   ├── README.md
│   └── project_requirements.txt
└── src
    ├── README.md
    ├── notebooks
    │   └── README.md
    ├── rules
    ├── scripts
    └── tests
```

# Arcus Project Template

## data/endpoints/

- Managed by study team
- Final results from research analysis
- Files generated to support papers or grants
- Organize in subdirectories if necessary




# Arcus Project Template

## data/ref-data/

- External or public datasets not supplied by Research IS or your lab, such as census data

RIS has recommended structure: https://github.research.chop.edu/RIS/reference_data


# Arcus Project Template

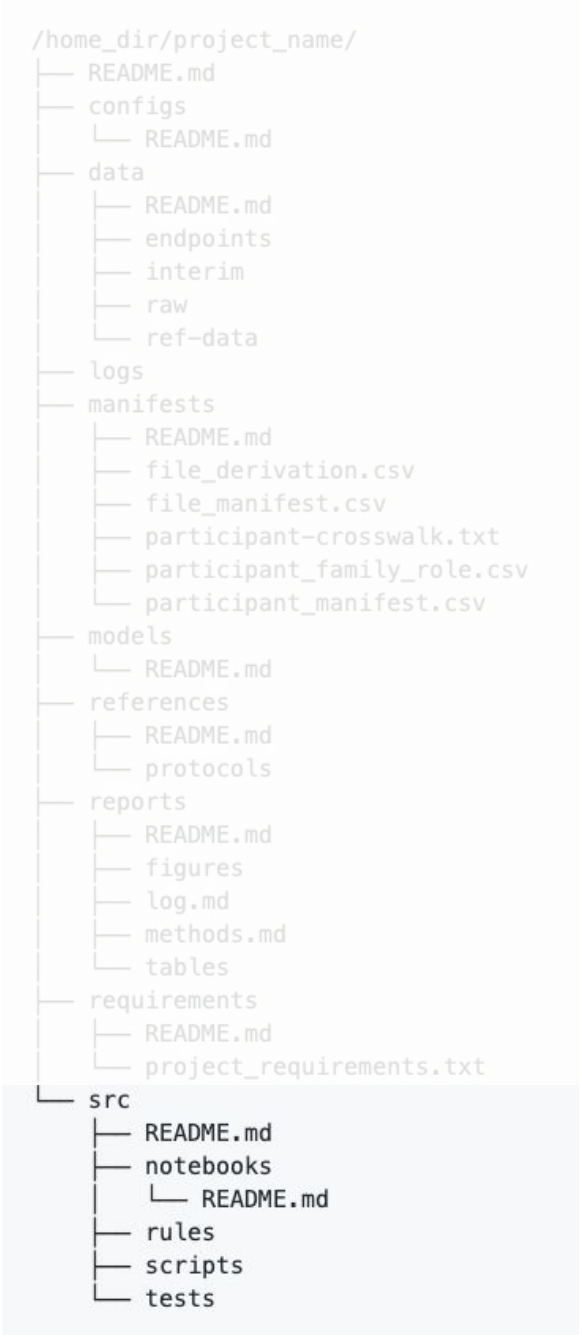
## manifests/

- Managed by Arcus
- Inventory of all raw & endpoint data
- Inventory of all participants and their related cohorts/samples/family roles, as applicable
- Associates IDs with all data files
- Documents relationships between files in pipelines/workflows

# Arcus Project Template

```
/home_dir/project_name/
├── README.md
├── configs
│   └── README.md
├── data
│   ├── README.md
│   ├── endpoints
│   ├── interim
│   ├── raw
│   └── ref-data
├── logs
├── manifests
│   ├── README.md
│   ├── file_derivation.csv
│   ├── file_manifest.csv
│   ├── participant-crosswalk.txt
│   ├── participant_family_role.csv
│   └── participant_manifest.csv
├── models
│   └── README.md
├── references
│   ├── README.md
│   └── protocols
├── reports
│   ├── README.md
│   ├── figures
│   ├── log.md
│   ├── methods.md
│   └── tables
├── requirements
│   ├── README.md
│   └── project_requirements.txt
└── src
    ├── README.md
    ├── notebooks
    │   └── README.md
    ├── rules
    ├── scripts
    └── tests
```

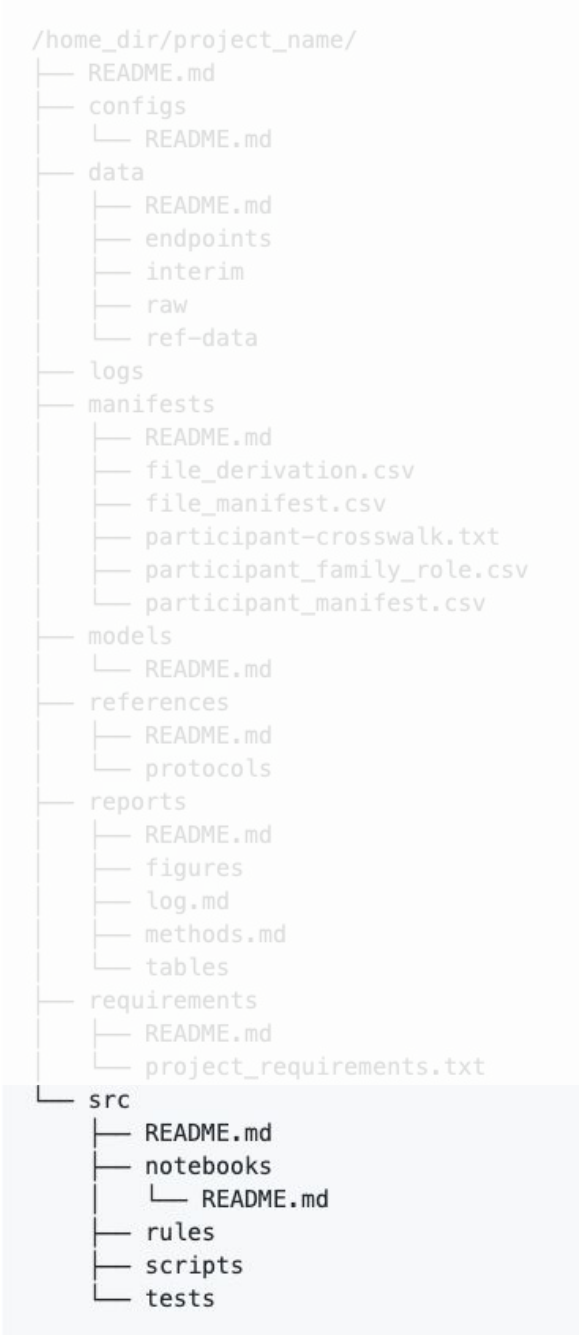
**Access tools**




# Arcus Project Template

## src/

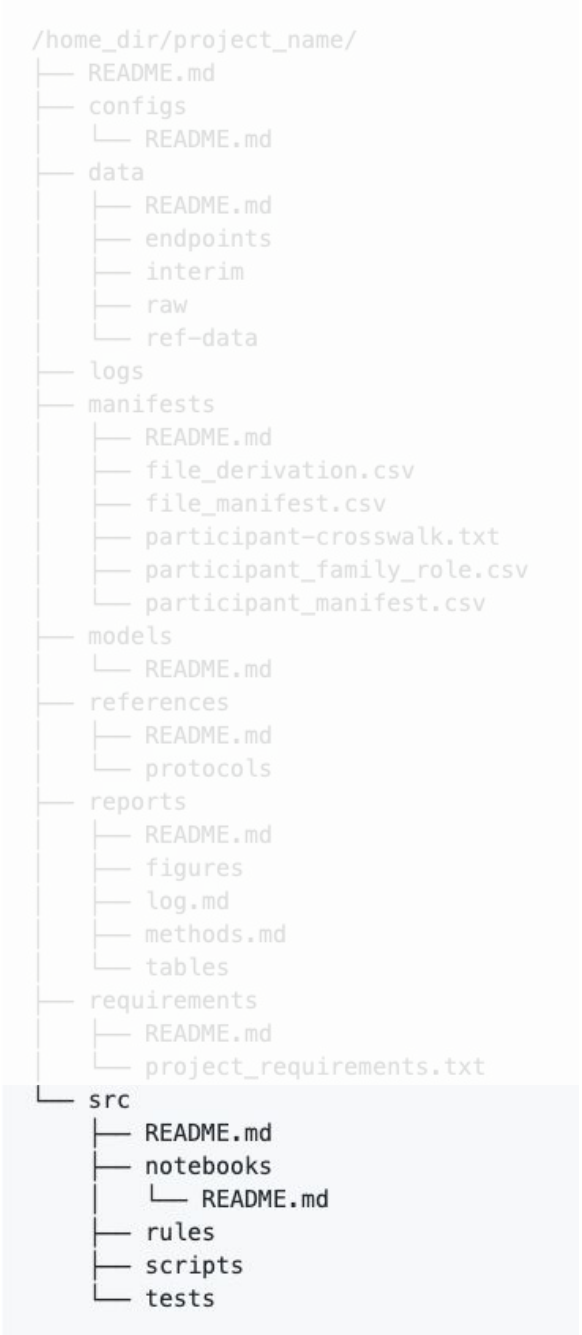
- Managed by study team
- Access tools required of the research data itself
- Version control is important!
- Subdirectory folders can be customized and added as needed




# Arcus Project Template

**src/** subfolders

- Common/useful options
  - notebooks/
    - Jupyter, Beaker, Zeppelin, etc.
  - scripts/
    - Custom software, code, scripts




# Arcus Project Template

**src/** subfolders

- Other examples of options
  - rules/
    - for computational workflows
  - test/
    - unit testing for code
  - customizable to team needs

```
/home_dir/project_name
├── README.md*
├── configs
├── data*
├── logs
├── manifests
├── models
├── references
├── reports
├── requirements
└── src*
```

# Arcus Project Template

* required components

```
/home_dir/project_name
├── README.md*
├── configs
├── data*
├── logs
├── manifests
├── models
├── references
├── reports
├── requirements
└── src*
```

# Arcus Project Template

* required components

- Additional folders can be added if useful
  - Research administration file
- Study teams can implement template outside of Arcus Lab environment to harmonize research management

# Arcus Research Data Management

Free consultations and additional resources:

- README structure
- Narrative descriptions of research
- Filenaming
- Data dictionaries
- Ontologies
- Data collection/processing
- Extract, transform, load (ETL)
- REDCap survey methodology
- NIH's data management plan requirements

email arcus.rdm@chop.edu