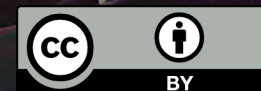


Managing data

**Why it matters,
when it is important,
and how to do it**

Nathan Cummings,
Samuel Jackson
UKAEA



Exercise

- Either on your own, or with the person next to you:
 - Find a dataset online (scientific data ideally)
 - Access the data
 - Plot something that shows that you know what the data mean
- Answer 3 questions about them:
 - Where are the data from?
 - Who/how/when/(why?)
 - How do you know what the data are/represent?
 - What are you allowed/not allowed to do with the data?
- 10 minutes (ish)... **GO!**

The logo features the word "FAIR" in a large, bold, white sans-serif font. Below it, the word "Principles" is written in a white, elegant script font. The background is a dark grey-blue gradient. On the left side, there are three vertical bars of increasing height and decreasing width, colored in shades of blue and teal.

FAIR

Principles

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al.
The FAIR Guiding Principles for scientific data
management and stewardship. Sci Data3, 160018
(2016). <https://doi.org/10.1038/sdata.2016.18>

F

Findable

A

Accessible

I

Interoperable

R

Reusable

Findable

- (Meta)data are assigned a globally unique and persistent identifier
- Data are described with rich metadata
- Metadata clearly and explicitly include the identifier of the data they describe
- (Meta)data are registered or indexed in a searchable resource

A I R

F Accessible

I R

- (Meta)data are retrievable by their identifier using a standardised communications protocol
 - The protocol is open, free, and universally implementable
 - The protocol allows for an authentication and authorisation procedure, where necessary
- Metadata are accessible, even when the data are no longer available

F A Interoperable R

- (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- (Meta)data use vocabularies that follow FAIR principles
- (Meta)data include qualified references to other (meta)data

F A I Reusable

- (Meta)data are richly described with a plurality of accurate and relevant attributes
 - (Meta)data are released with a clear and accessible data usage license
 - (Meta)data are associated with detailed provenance
 - (Meta)data meet domain-relevant community standards

**Experimental
data and FAIR
– why should
you care?**

Research is easier when you're
FAIR

Findable – **easily locate** the data
needed for training sets etc...

Accessible – fosters **collaboration**

Interoperable – **easily** work with
different workflows & analysis tools

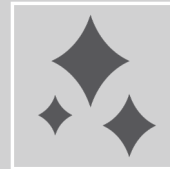
Reusable – descriptive metadata
provides context



When does it matter?



More often than you might think!



Any time you're creating new
data



When given data that isn't FAIR,
consider improving the
FAIRness (as long as you are
allowed to!)

What else can you do?

When developing software tools:

- As open as possible
- Licence
- DOCUMENT

When using data:

- Assess the **FAIR**ness
- Feedback to producers
- Recommend to others

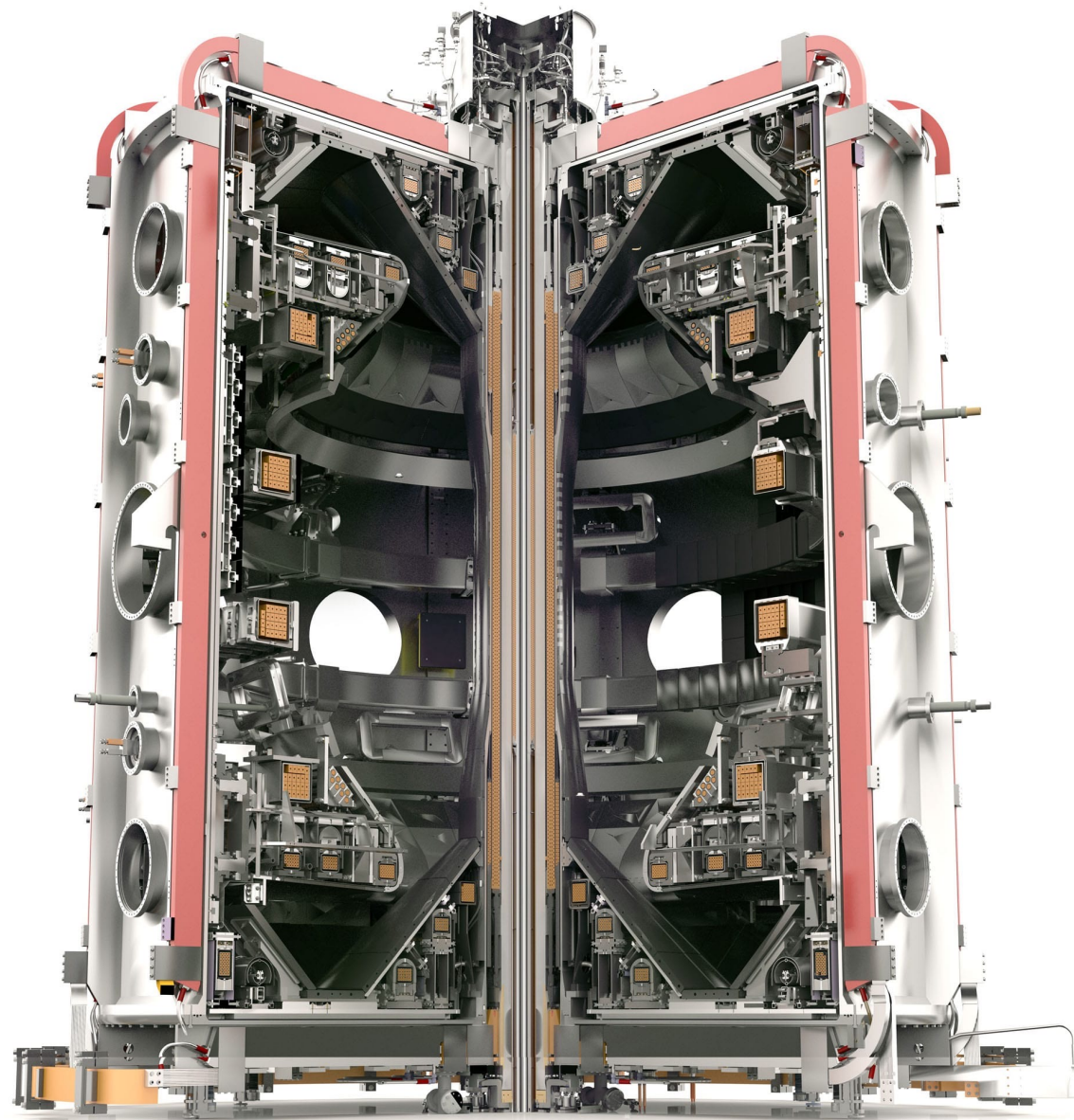
What else can you do?

When producing data:

- Use **standards** where possible
 - **ISO (International Organization for Standardization)**: They provide various standards, e.g., ISO 19115 for geographic information
 - **W3C (World Wide Web Consortium)**: They offer standards like RDF (Resource Description Framework) and DCAT (Data Catalog Vocabulary)

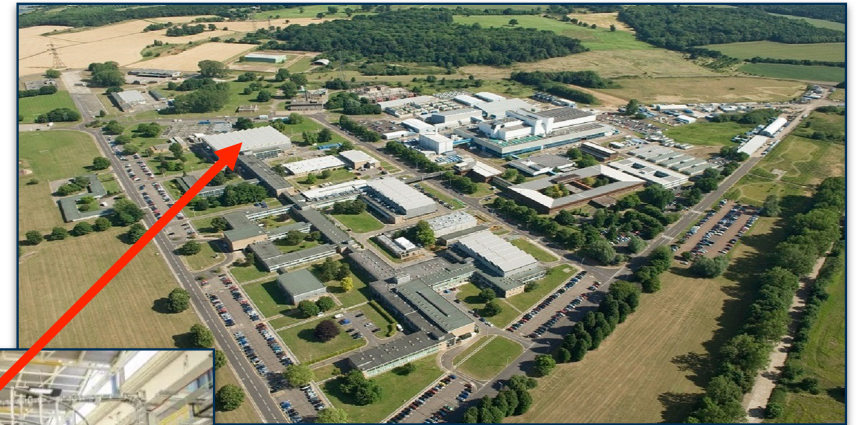
Ask!

- Most organisations have policies for data management, including ways to store/distribute your data

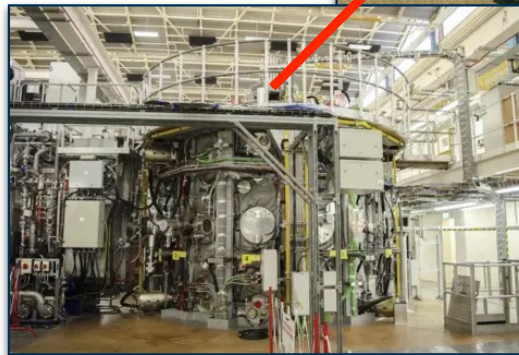


MAST

- MAST (Mega Amp Spherical Tokamak)
- Spherical tokamak design commissioned by EURATOM/UKAEA
- Built at Culham Centre for Fusion Energy, Oxfordshire, UK
- Experiments ran from 1999 through to 2013
- Produced ~30,000 shots over its history
- Succeeded by MAST Upgrade (MAST-U) in 2020



Culham Centre for Fusion Energy, UK



MAST Tokamak

Motivation

We want to:

- Have software tools that are robust and can scale
- Gain expertise from complementary domains
- Collaborate with the wider world
 - Fusion energy, Data, and AI/ML communities

We need:

- Open access with minimal barriers.
- Integrate with data analysis & reduction tools that scale.
- Integrate with domain agnostic tools.
 - We cannot afford to build everything ourselves.
- Perform search, retrieval, and analysis across the historical record



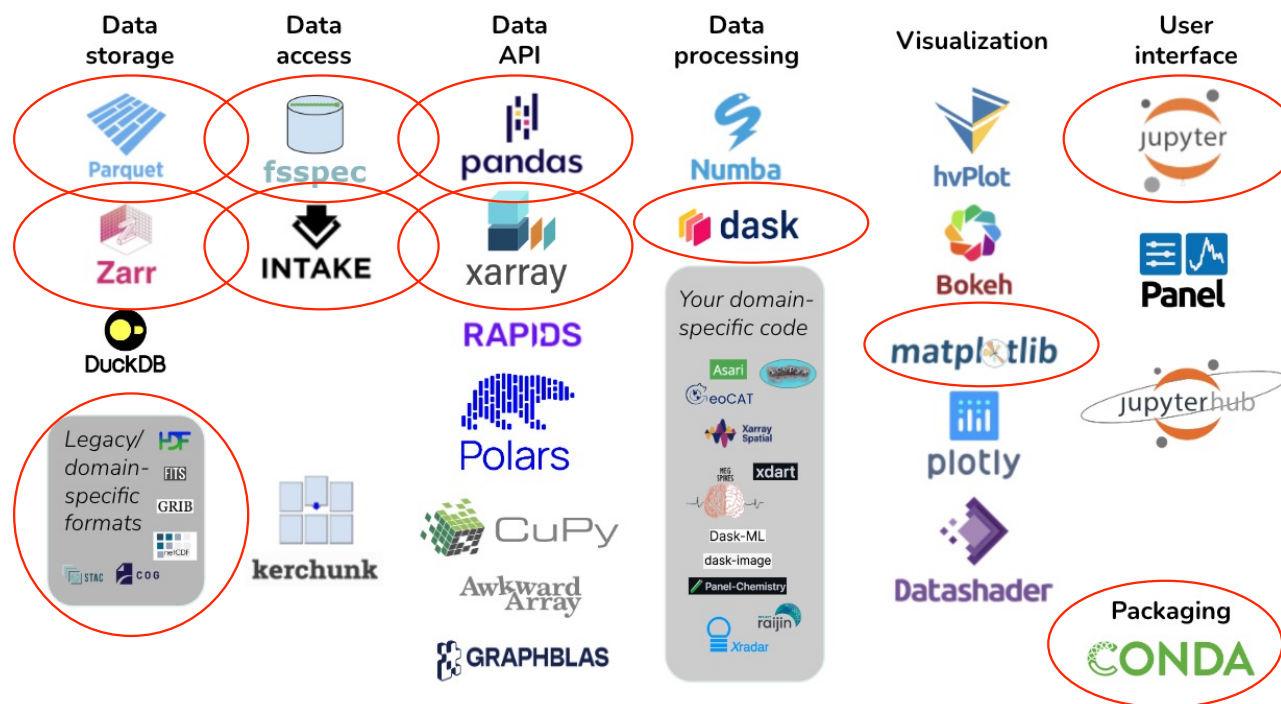
Project Objectives

Goal: “To produce a framework for public access to MAST data in a FAIR (Findable, Accessible, Interoperable, and Reusable) manner”.

- Data must be easily **findable** through the metadata
- Data must be exposed in an **interoperable** format
- Prioritize **performance optimization** for artificial intelligence (AI) and machine learning (ML) workflows
- Minimize **loading** and **transferring** data (lazy loading)
- Support **data analysis** and **ML/AI** frameworks
- Support **larger-than-memory** & **parallel** computation
- Be **publicly** accessible

Pandata Stack

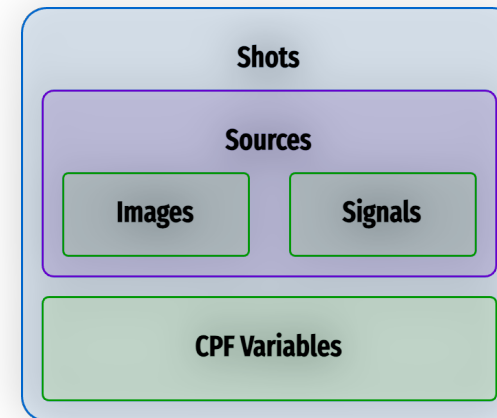
Pandata stack is an **open-source** set of **interoperable**, **composable**, and **domain agnostic** software technologies for data analysis and scientific computation.



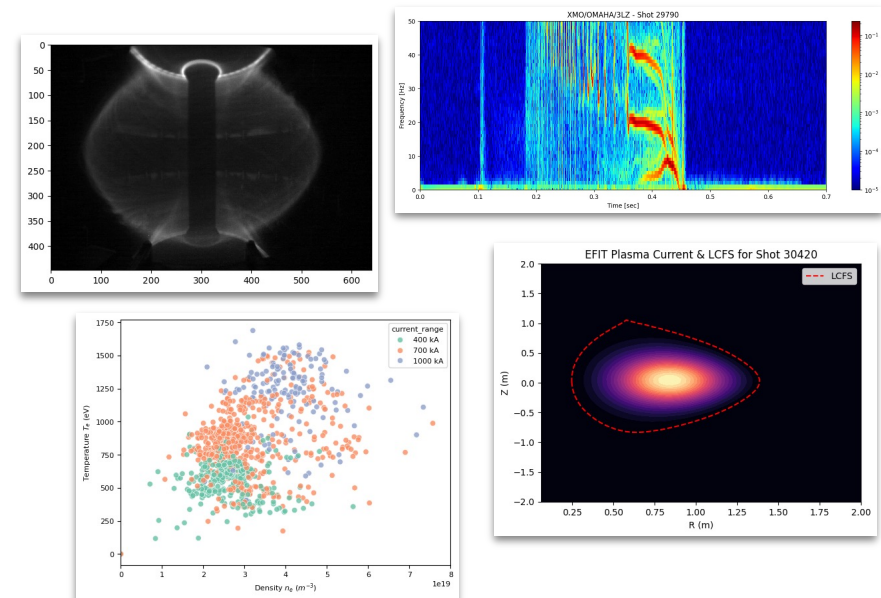
MAST Diagnostic Data

MAST Data can be thought of in terms of:

- **Shots:** A single experimental shot taken by the machine.
- **Sources:** Each shot contains multiple diagnostic sources.
 - Examples include: Mirnov Coils, Thompson scattering, EFIT output etc.
- **Signals:** Each source contains multiple recorded quantities.
 - In MAST these were conceptually split into “signals” and “images”.
- **Summary Physics Variables:** Additional summary statistics documenting a shot.
 - e.g. max plasma current, beta, confinement time



Conceptual overview of different types of data from MAST



Indexing

Our metadatabase indexes the data records within each file.

We index on three levels:

- Shots
- Signals
- Sources

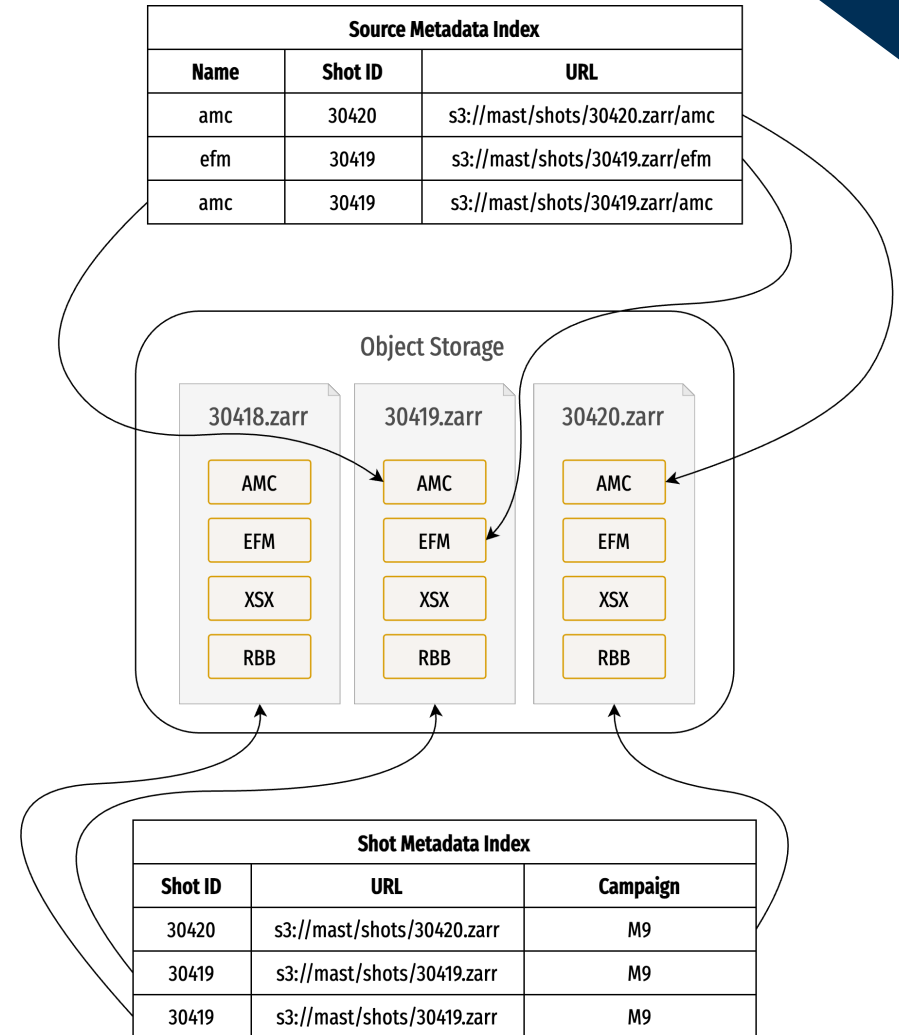
Each item has a UUID assigned to it and references a URL which links to the object storage.

Database implemented with PostgreSQL

FAIR Principles

F4. (Meta)data are registered or indexed in a searchable resource

A2. Metadata are accessible, even when the data are no longer available



Metadata APIs: REST

The image shows a screenshot of the MAST Archive (0.1.0) REST API documentation. On the left, a sidebar lists various API endpoints such as 'Get Shots', 'Get Shots Aggregate', 'Get Shot', 'Get Signals For Shot', 'Get Signals', 'Get Signals Aggregate', 'Get Signal', 'Get Shot For Signal', 'Get Cpf Summary', 'Get Scenarios', 'Get Sources', 'Get Sources Aggregate', 'Get Single Source', 'Get Signals Stream', 'Get Shots Stream', and 'Get Source Stream'. The main area displays the 'Get Shots' endpoint, which provides information about experimental shots. It includes a 'Download OpenAPI specification' button and a 'QUERY PARAMETERS' section with fields like 'fields', 'filters', 'sort', 'page', and 'per_page'. A red arrow points from the 'filters' parameter in the documentation to a 'REST API query result' window. This window shows a GET request to '/json/shots' with a response sample in JSON format, including fields like 'shot_id', 'uuid', 'url', 'timestamp', 'pre-shot_description', 'post-shot_description', 'campaign', 'reference_shot', 'scenario', 'heating', 'pellets', 'rmp_coil', 'current_range', 'divertor_config', 'plasma_shape', 'comissioner', and 'facility'.

REST API query result

REST API implemented with fastapi,
sqlmodel, and sqlalchemy

Experimented with GraphQL written on top with
strawberry

The image shows a screenshot of a GraphQL query explorer. On the left, a query is written in a monospace font:

```
1 query {
2   all_shots (limit: 10, where: {shot_id: {lt: 30420}}) {
3     shots {
4       shot_id
5       divertor_config
6       heating
7       timestamp
8     }
9   }
10 }
```

 On the right, the query result is displayed in a JSON-like format, showing a list of shots with their respective attributes.

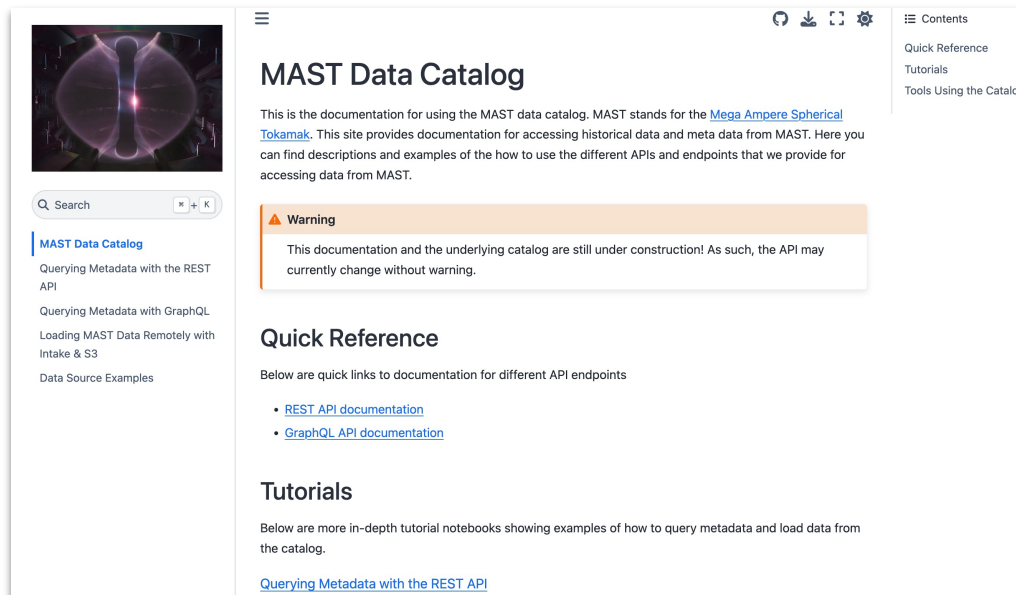
GraphQL query explorer

Summary

We developed a data infrastructure solution for the history of the MAST experiment

We provide a public REST API for the metadata

We provide a public the history of the MAST data in cloud object storage



Test site: <https://mastapp.site/>

