Managing data

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

Nathan Cummings,
Samuel Jackson
UKAEA

UK Atomic Energy Authority

Exercise

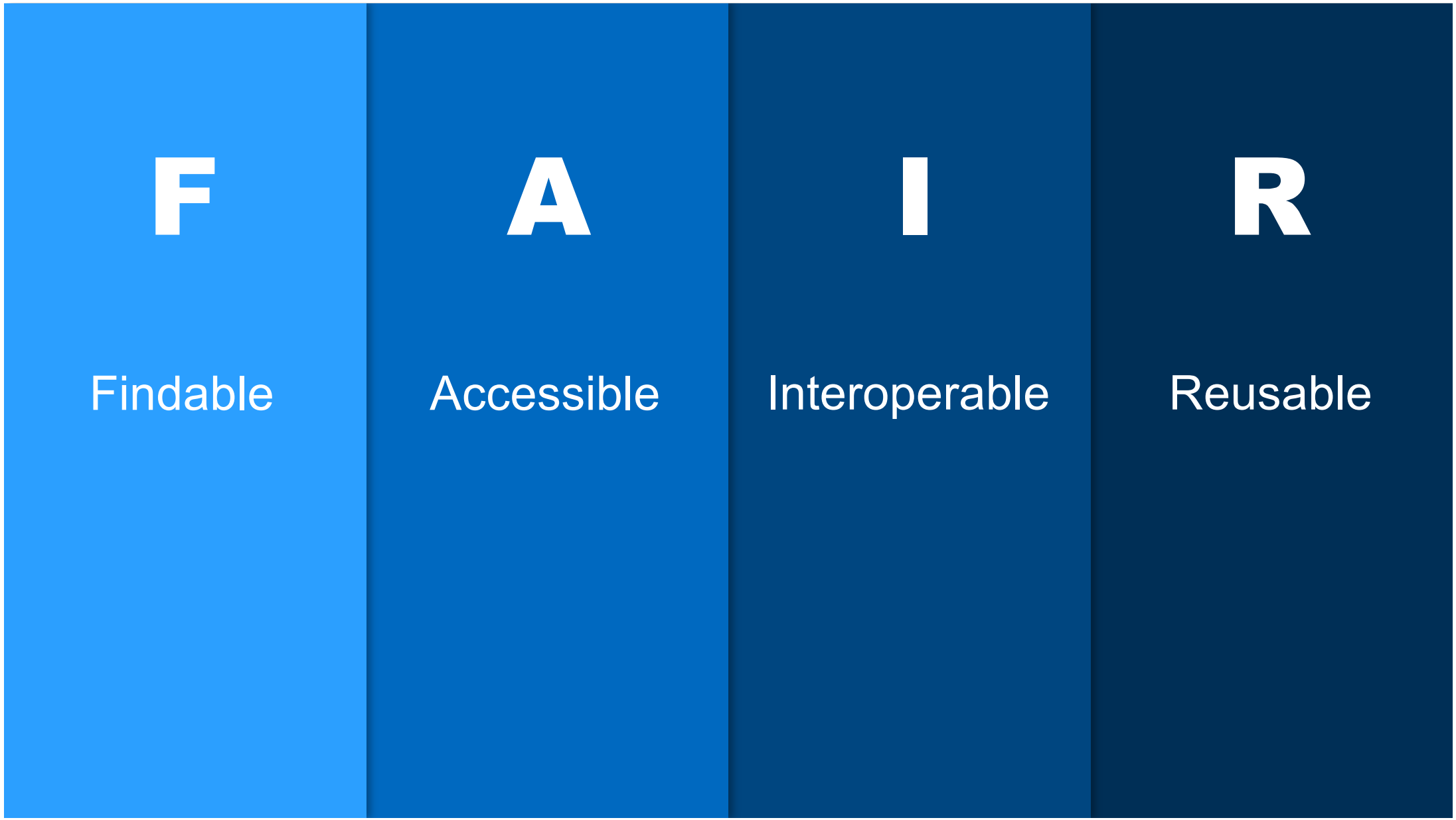
- Either on your own, or with the person next to you:
  - Find a dataset online (scientific data ideally)
  - Access the data
  - Plot something that shows that you know what the data mean

- 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!**

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

A I R

- (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

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 R

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

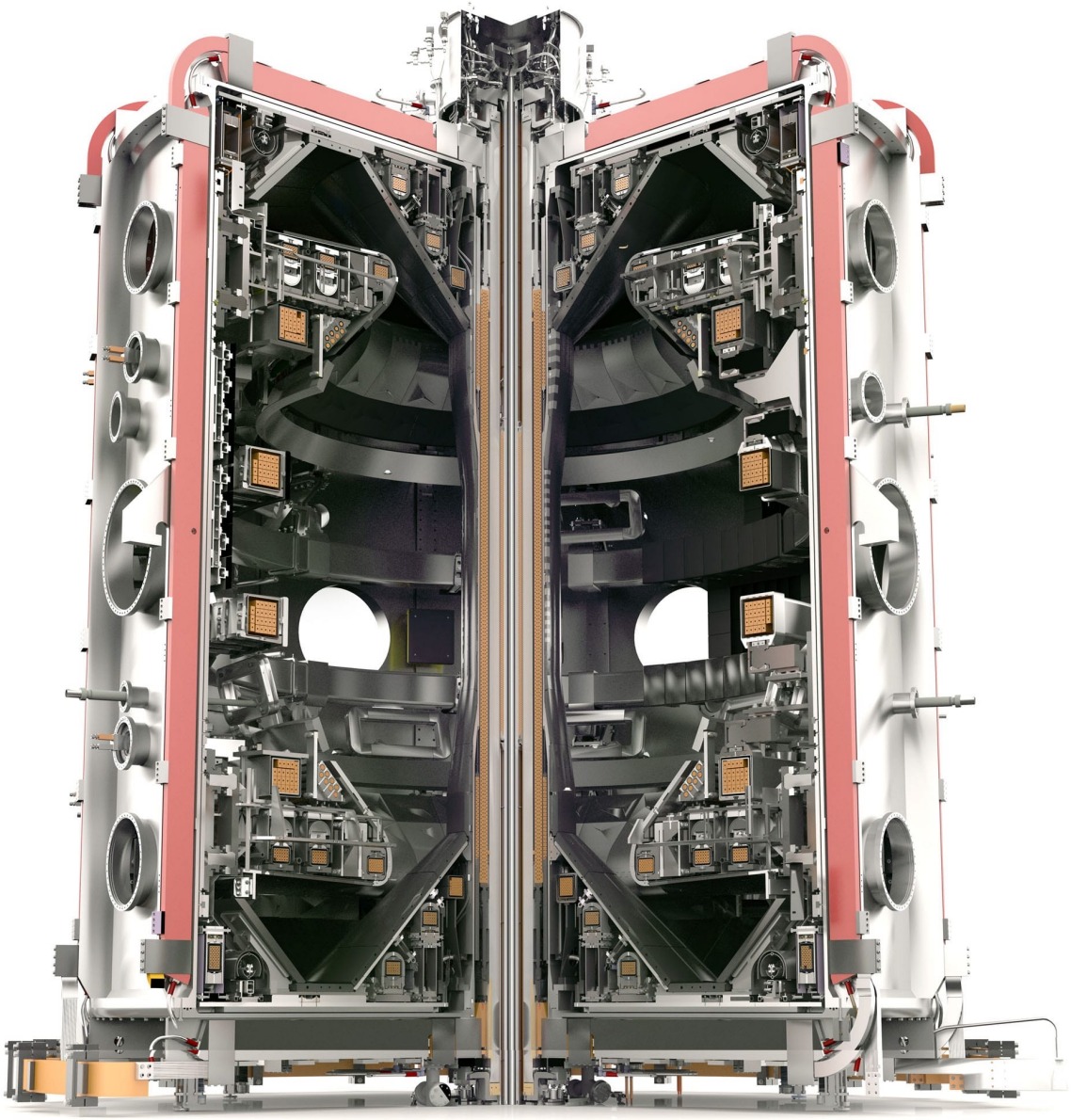
What else can you do?

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)

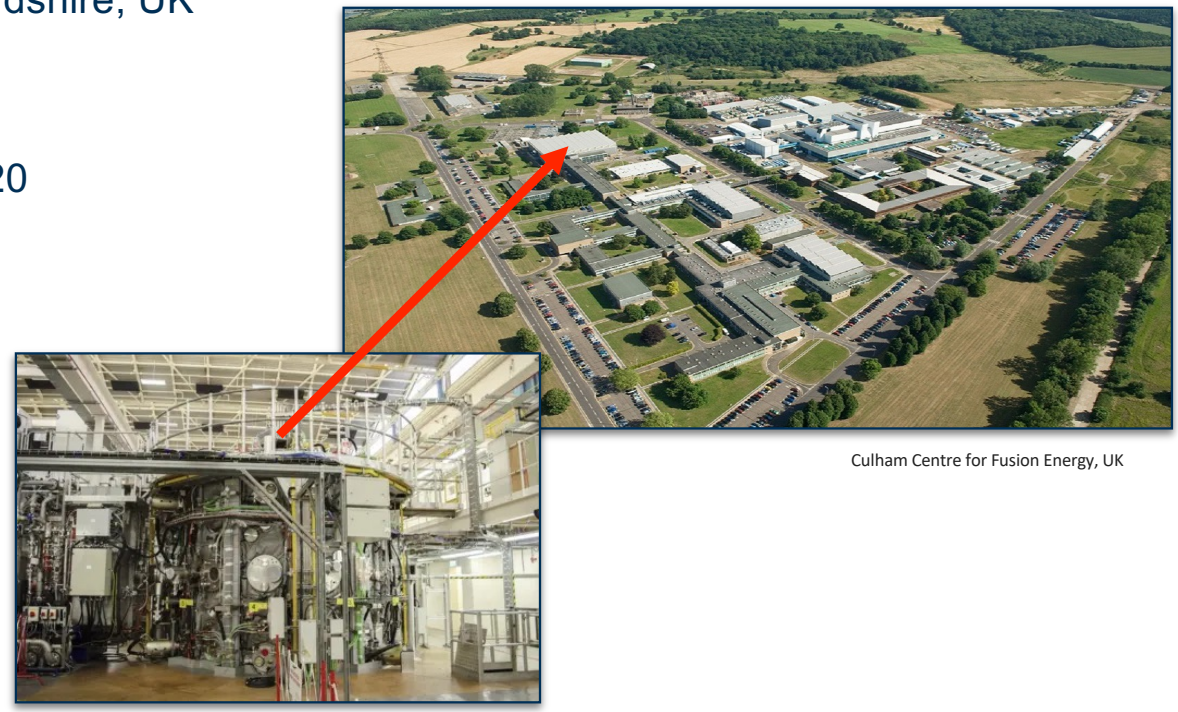
Ask!

- 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

UK Atomic Energy Authority

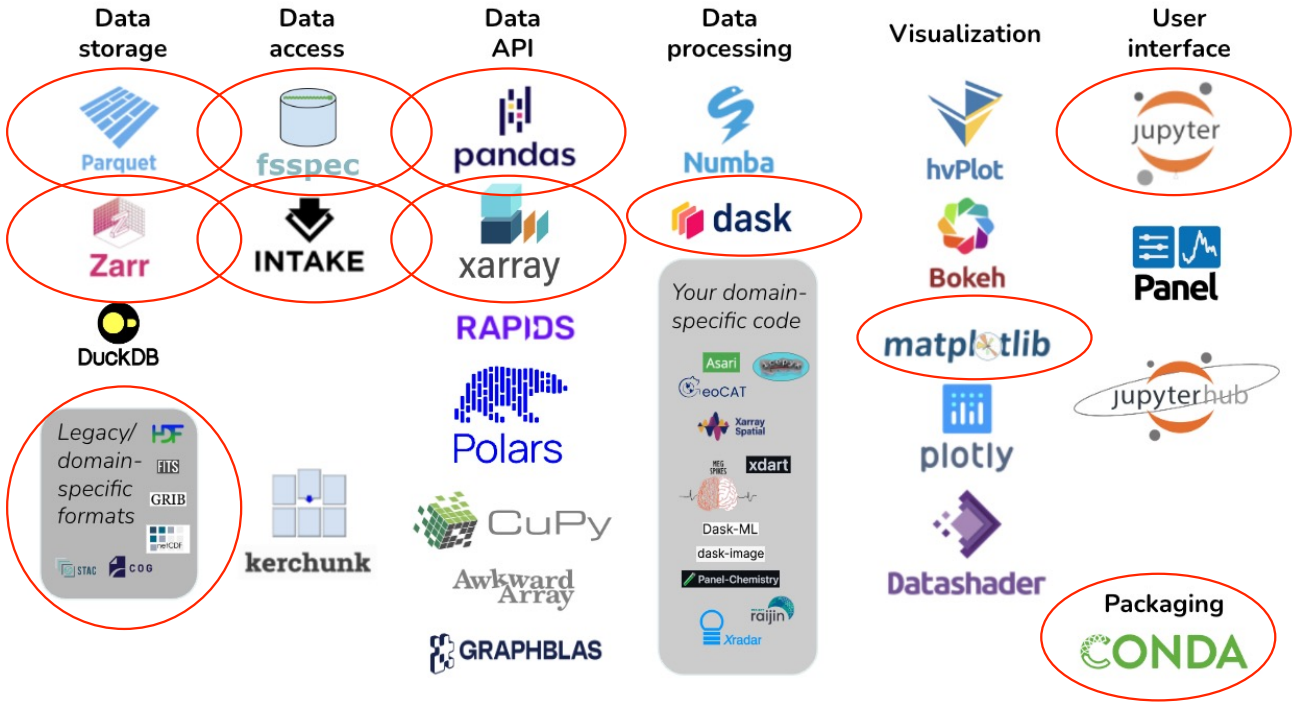
Goal: "*To produce a framework for public access to MAST data in a FAIR (Findable, Accessible, Interoperable, and Reusable) manner*".

- Data must be easily **findable** through the metadata
- Data must be in 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

UK Atomic Energy Authority

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

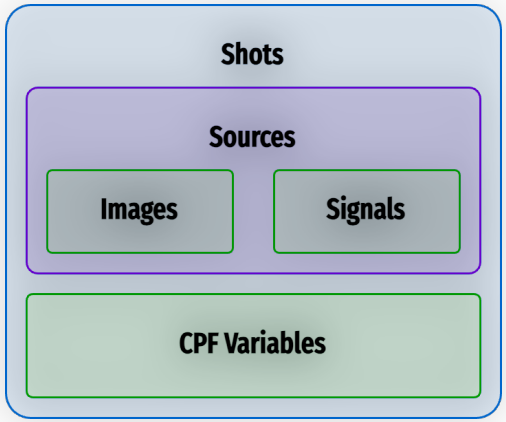


Bednar J.A., Durant M. The Pandata Scalable Open-Source Analysis Stack.

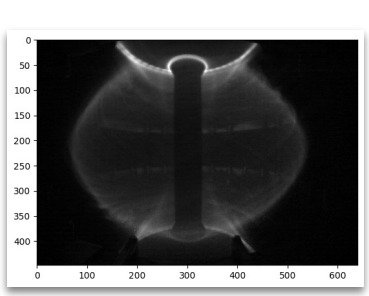
MAST Diagnostic Data

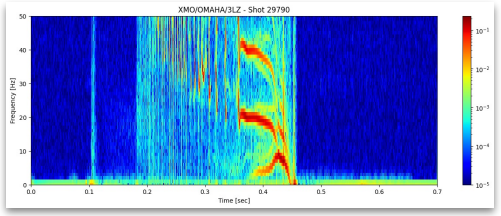
MAST Data can be thought of in terms of:

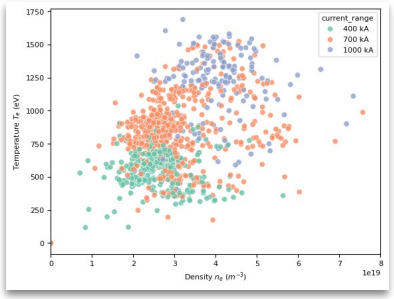
- **Shots:** A single experimental shot taken by the machine.
- **Sources:** Each shot contains multiple diagnostic sources.
  - Examples include: Mirnov Coils, Thompson scattering, EFIT output etc.
- **Signals:** Each source contains multiple recorded quantities.
  - In MAST these were conceptually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

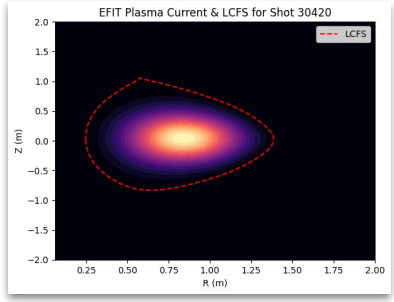


Conceptual overview of different types of data from MAST









Indexing

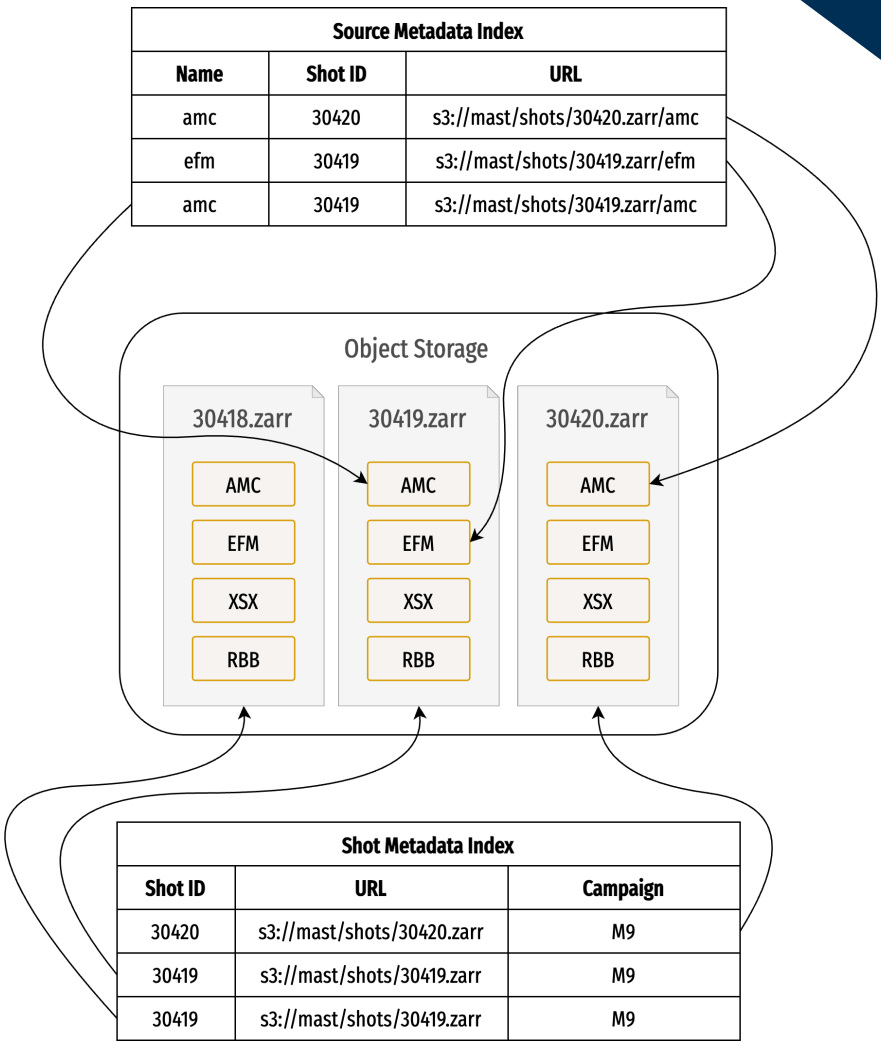
Our metadatabase indexes the data records within each file.

We index on three levels:

- Shots
- Signals
- Sources

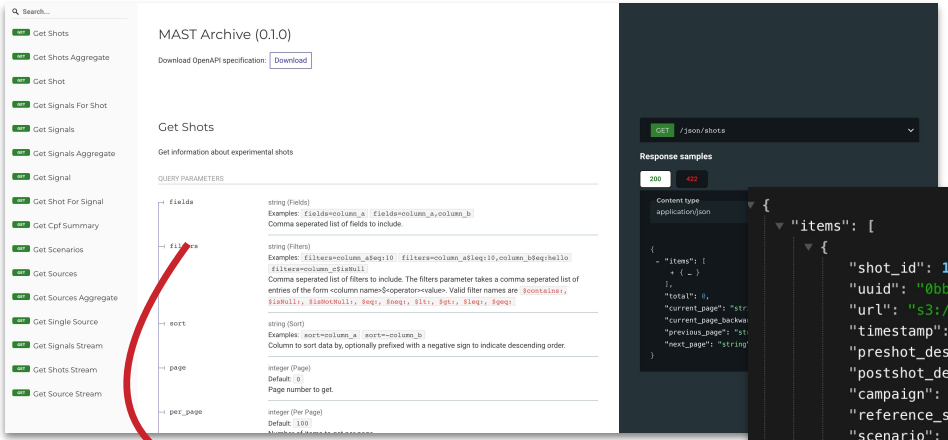
Each item has a UUID assigned to it and references a URL which links to the object storage.

Database implemented with PostgreSQL



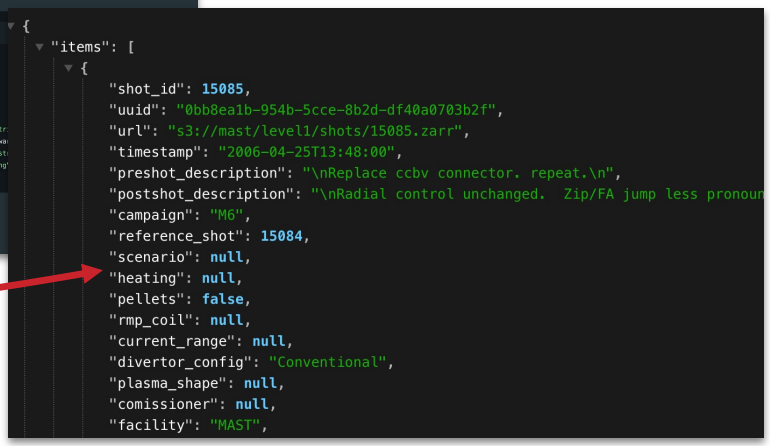
FAIR Principles
F4. (Meta)data are registered or indexed in a searchable resource
A2. Metadata are accessible, even when the data are no longer available

Metadata APIs: REST



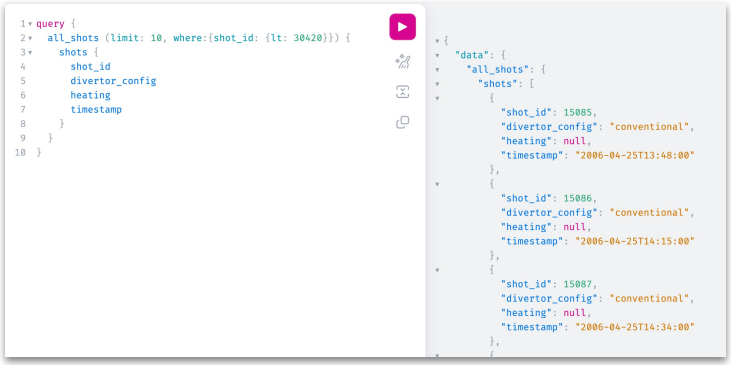
REST API Documentation

REST API query result



REST API implemented with `fastapi`, `sqlmodel`, and `sqlalchemy`

Experimented with GraphQL written on top with `strawberry`



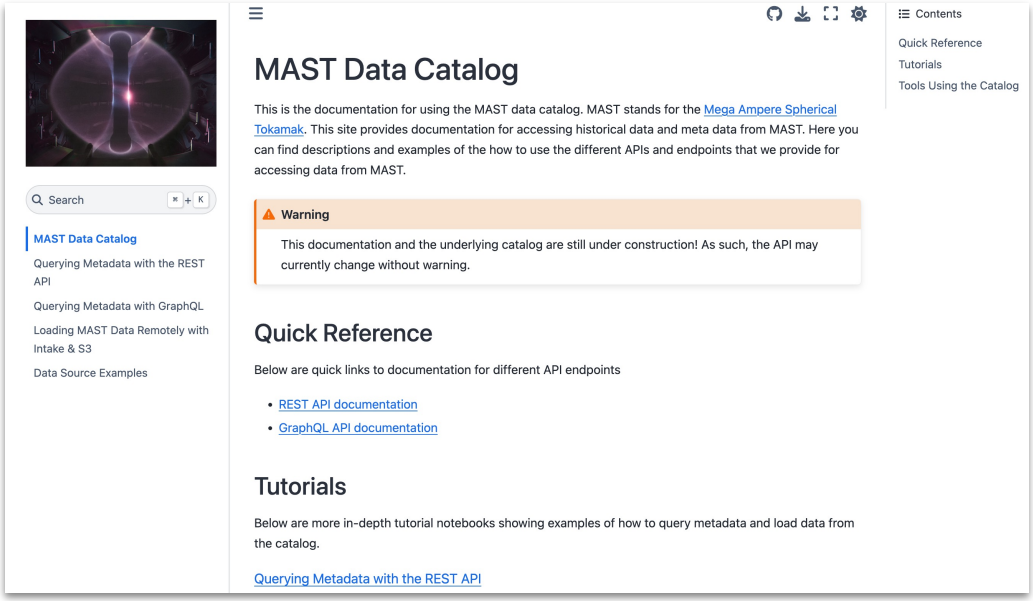
GraphQL query explorer

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/