

The IAEA Fusion Data Lake Project

**Accelerating AI and Big Data Applications
through Open Science and FAIR Data**

D.S. Gahle, M. Barbarino, *et al*, IAEA
6th IAEA Technical Meeting on Fusion Data
Processing, Validation and Analysis
September 9th-12th, Fudan University, Shanghai

AI for Fusion and Fusion Data Lake Collaboration

****IAEA AI for Fusion CRP Partners**

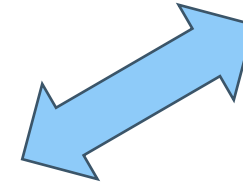
The Australian National University (*Australia*), Institute of Plasma Physics, Chinese Academy of Sciences (*China*), Shanghai Jiao Tong University (*China*), Southwestern Institute of Physics (*China*), Huazhong University of Science and Technology (*China*), Southwestern Institute of Physics (*China*), Institute for Plasma Research (*India*), Eni S.p.A. (*Italy*), University of Cagliari (*Italy*), Osaka University (*Japan*), National Institute for Fusion Science (*Japan*), Korea Institute of Fusion Energy (*Korea*), Chung-Ang University (*Korea*), Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (*Spain*), Chalmers University of Technology (*Sweden*), Swiss Federal Institute of Technology Lausanne (*Switzerland*), Imperial College London (*UK*), Culham Centre for Fusion Energy – UK Atomic Energy Agency Culham Science Centre (*UK*), First Light Fusion Ltd. (*UK*), Massachusetts Institute of Technology (*USA*) (*USA*), Princeton Plasma Physics Laboratory (*USA*), University of Notre Dame (*USA*), General Atomics (*USA*), University of Wisconsin-Madison (*USA*), and College of William & Mary (*USA*).



International Atomic Energy Agency

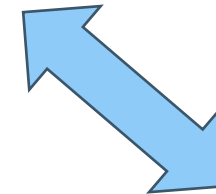


Fusion team
Project Owner
NAPC-Physics



MTIT BSS team
Business Software
Solutions

Webapp



MTIT CDO team
Chief Data Office

Data-backend

AI for Fusion CRP and the Fusion Data Lake Project

"AI for Fusion: accelerating fusion R&D with AI, through the creation of a platform and cross-community network for innovation and partnership" - AI for Fusion ([Homepage](#))

Who and what do we need for AI accelerated development of fusion energy?

The who:

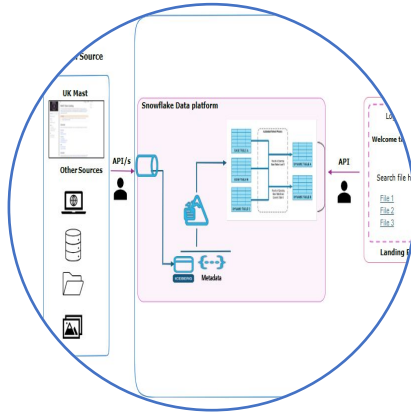
- Physicists
- Software developers
- Statisticians

The what:

- Surrogate models
- Digital twins
- Large scale data access!

The Fusion Data Lake: Accelerating AI development through Open Science and FAIR Data

Today's Topics



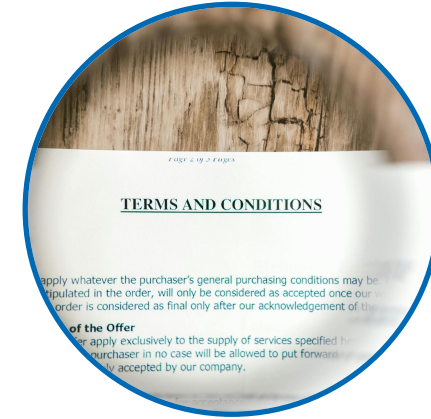
Platform Design

Architecture design, data and metadata models, and technology and data transfers



Proof of Concept

What has been developed, and what is coming next



Data Governance

Data access and terms of service



Platform Design: Goal

The Fusion Data Lake has three core deliverables as its goals:

- 1. Catalogue** – Query-able, central catalogue of experimental fusion experiments (devices, shots, and diagnostic signals)
- 2. Data Federation** – Decentralised network of experimental repositories accessible through a single point
- 3. Medium Term Storage** – Centralised storage for data projects

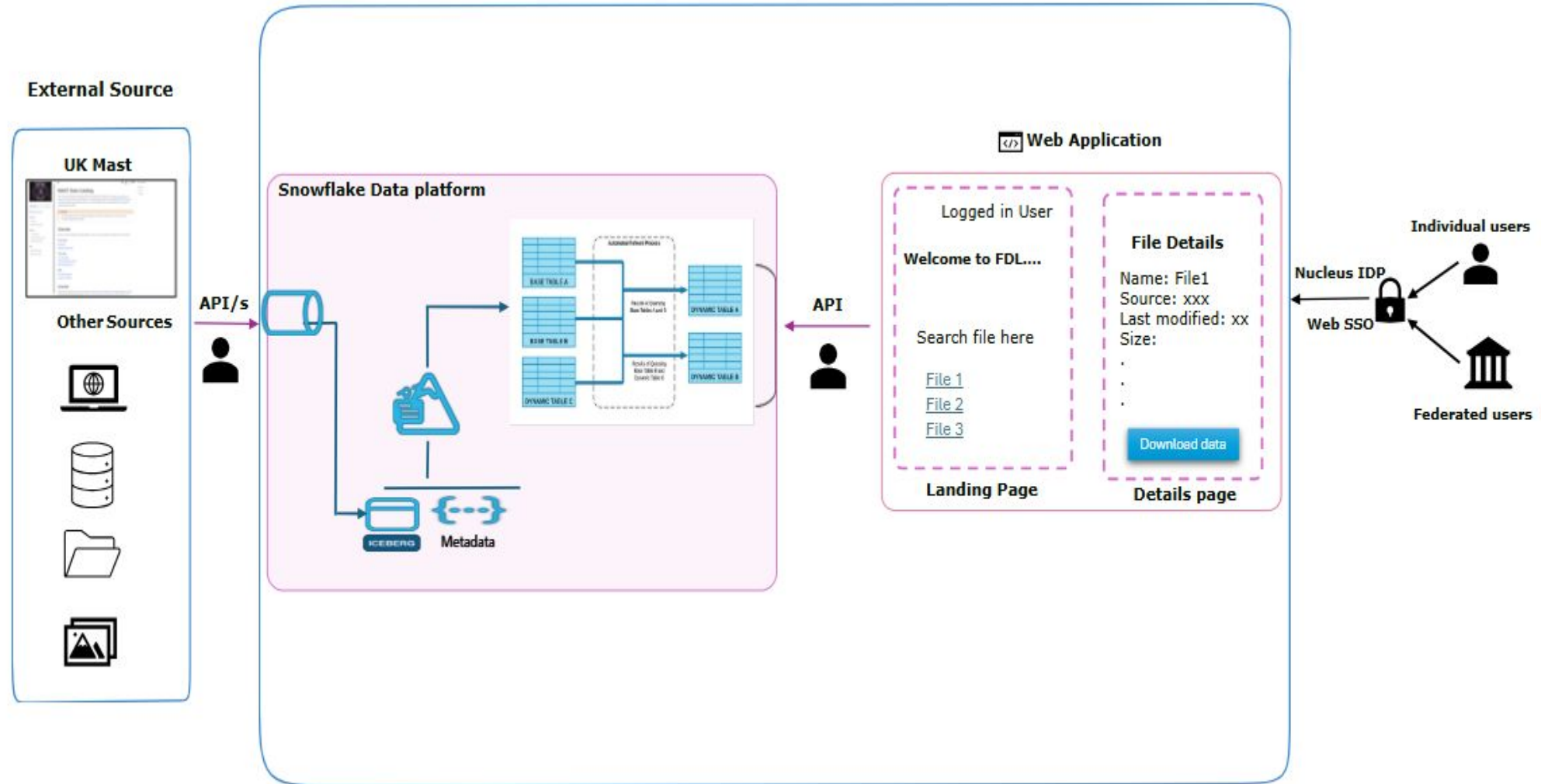
Platform Design: Goal

The Fusion Data Lake has three core deliverables as its goals:

1. **Catalogue** – Query-able, central catalogue of experimental fusion experiments (devices, shots, and diagnostic signals)
2. **Data Federation** – Decentralised network of experimental repositories accessible through a single point
3. **Medium Term Storage** – Centralised storage for data projects

Building an inclusive fusion data platform for everyone!

Platform Design: Architecture



Platform Design: Architecture

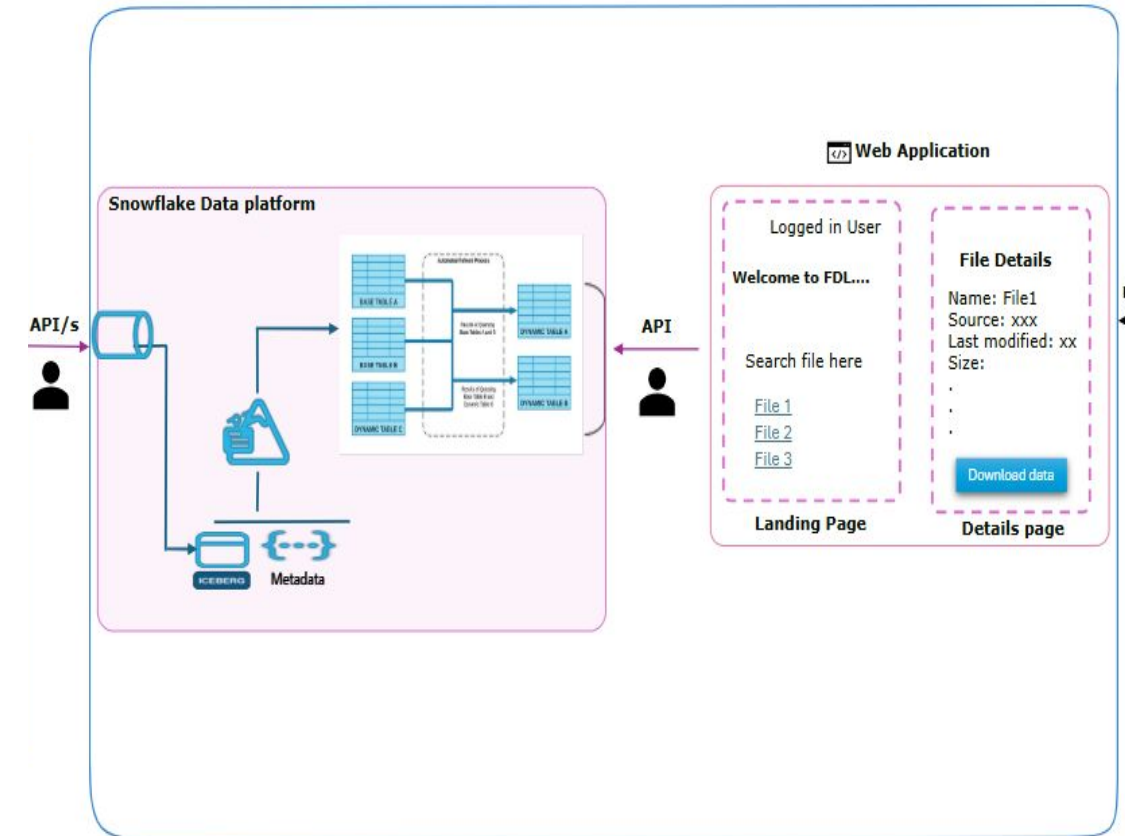
Tech stack:

- Compute - Snowflake
- Storage – Azure
- Webapp – C#/.Net

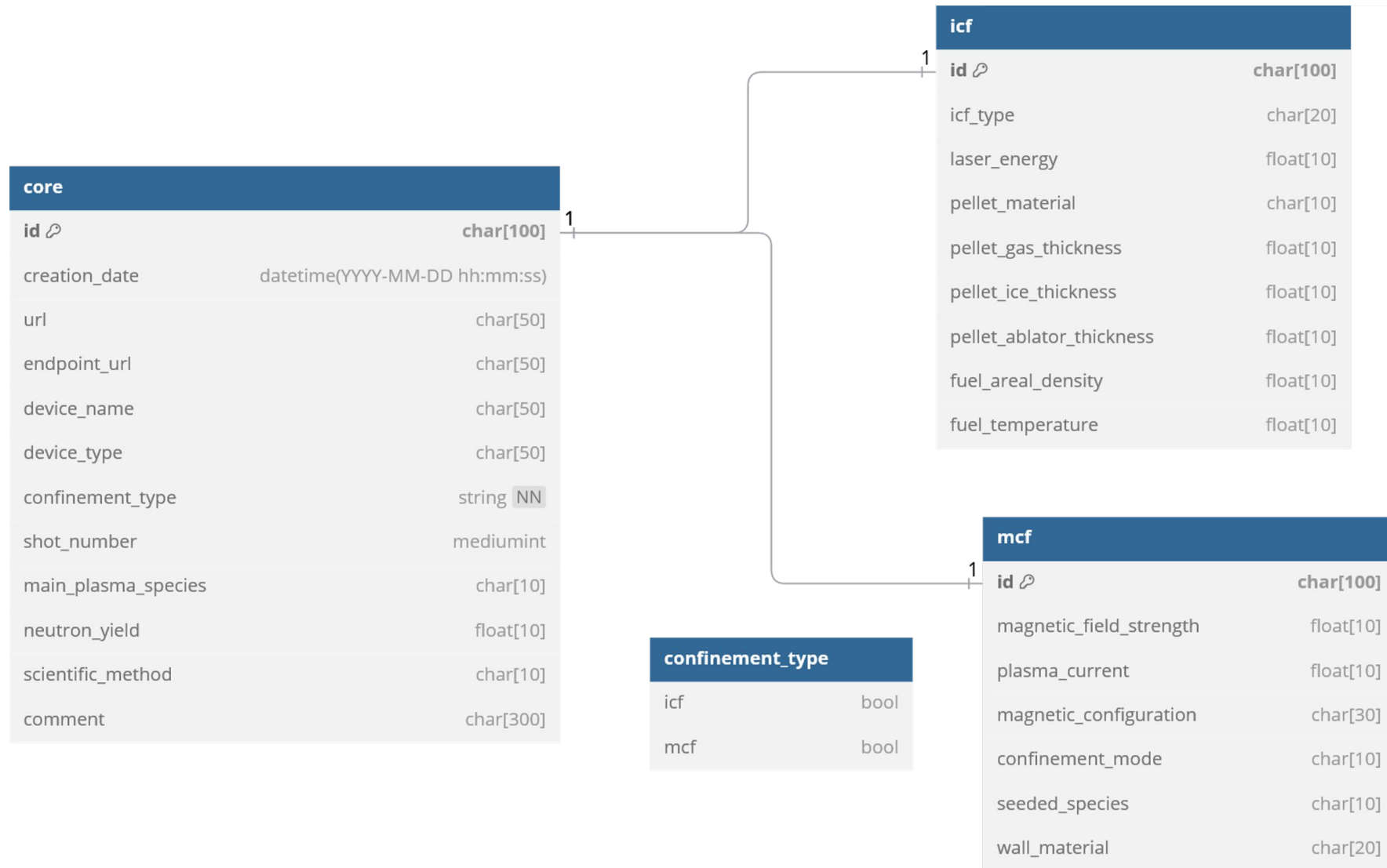
Data Engineering:

- Python (Snowflake native packages)
- Metadata driven ETL pipelines
 - Ingestions configuration
 - Transformation configuration
- Medallion structure

Data Engineering approach: Azure DevOps



Platform Design: Data and Metadata Model



View interactively here: [Fusion Data Lake - Minimal Metadata Model - dbdiagram.io](https://dbdiagram.io/Fusion%20Data%20Lake%20-%20Minimal%20Metadata%20Model)

Platform Design: Data & Metadata Model

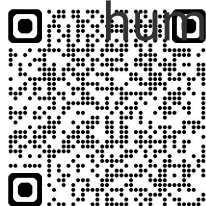
For a catalogue:

- Minimum data set for querying (IFE/MFE)

In general:

- Ontologies from the "ITER Data Dictionary" and global standards such as "Dublin Core"
- Units and data scales
- Data types and limits

Storage: Dynamic allocation (machine and human friendly)



View interactively here: [Fusion Data Lake - Minimal Metadata Model - dbdiagram.io](https://dbdiagram.io)

Platform Design: Technology and Data Transfers

Provisional Development Partners



Provisional Federation Partners

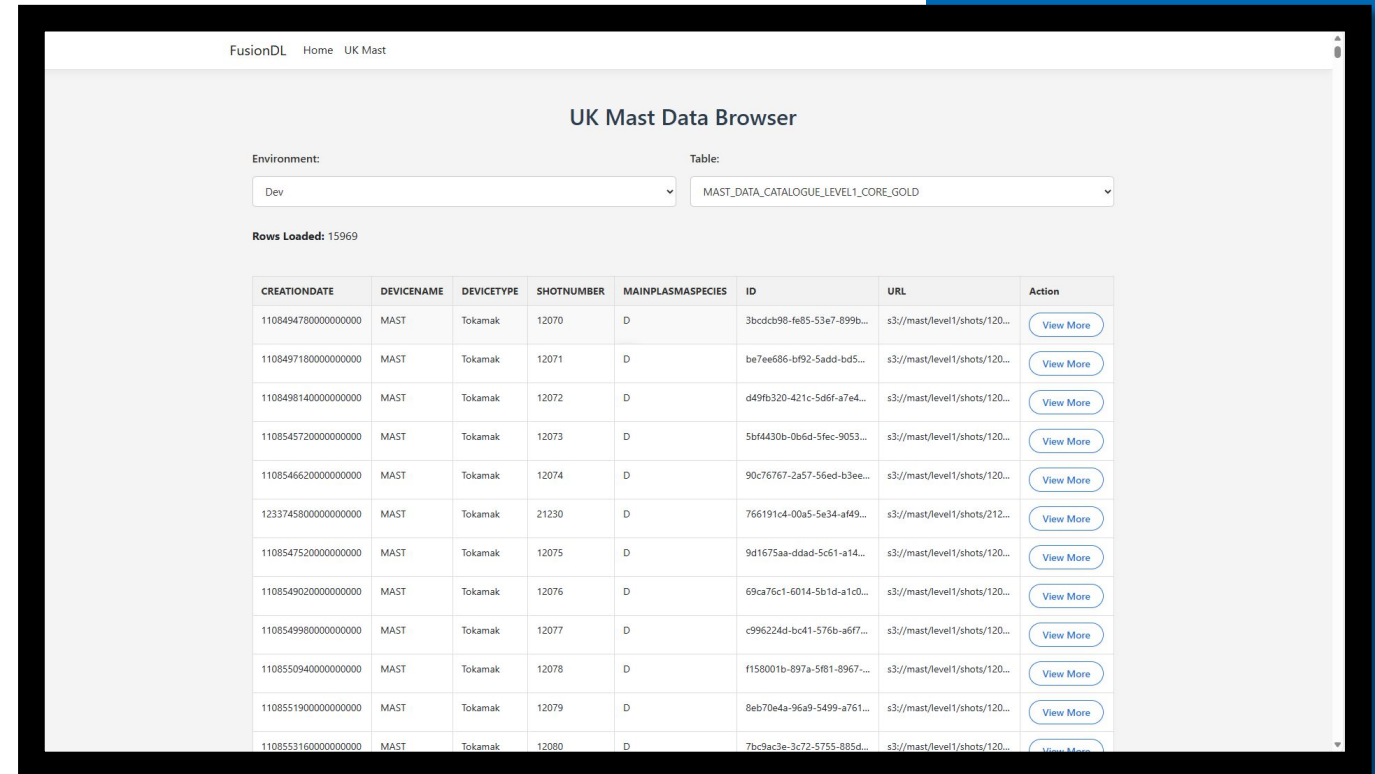


Do you want to share data and technology with the Fusion Data Lake project?

Proof of Concept: Phase 1 – Complete

What was done:

- Data pipeline
- MAST Data Catalog connection
- Medallion data storage
- Webapp catalogue



The screenshot shows a web application titled "UK Mast Data Browser". At the top, there are navigation links for "FusionDL", "Home", and "UK Mast". Below the title, there are two dropdown menus: "Environment:" set to "Dev" and "Table:" set to "MAST_DATA_CATALOGUE_LEVEL1_CORE_GOLD". A status message indicates "Rows Loaded: 15969". The main content is a table with the following columns: CREATIONDATE, DEVICENAME, DEVICETYPE, SHOTNUMBER, MAINPLASMASPECIES, ID, URL, and Action. The table contains 12 rows of data, each with a "View More" button in the Action column.

CREATIONDATE	DEVICENAME	DEVICETYPE	SHOTNUMBER	MAINPLASMASPECIES	ID	URL	Action
11084947800000000000	MAST	Tokamak	12070	D	3bcdcb98-fe85-53e7-899b...	s3://mast/level1/shots/120...	View More
11084971800000000000	MAST	Tokamak	12071	D	be7ee686-bf92-5add-bd5...	s3://mast/level1/shots/120...	View More
11084981400000000000	MAST	Tokamak	12072	D	d49fb320-421c-5d6f-a7e4...	s3://mast/level1/shots/120...	View More
11085457200000000000	MAST	Tokamak	12073	D	5bf4430b-0b6d-5fec-9053...	s3://mast/level1/shots/120...	View More
11085466200000000000	MAST	Tokamak	12074	D	90c76767-2a57-56ed-b3ee...	s3://mast/level1/shots/120...	View More
12337458000000000000	MAST	Tokamak	21230	D	766191c4-00a5-5e34-af49...	s3://mast/level1/shots/212...	View More
11085475200000000000	MAST	Tokamak	12075	D	9d1675aa-ddad-5c61-a14...	s3://mast/level1/shots/120...	View More
11085490200000000000	MAST	Tokamak	12076	D	69ca76c1-6014-5b1d-a1c0...	s3://mast/level1/shots/120...	View More
11085499800000000000	MAST	Tokamak	12077	D	c996224d-bc41-576b-a6f7...	s3://mast/level1/shots/120...	View More
11085509400000000000	MAST	Tokamak	12078	D	f158001b-b97a-5f81-8967...	s3://mast/level1/shots/120...	View More
11085519000000000000	MAST	Tokamak	12079	D	8eb70e4a-96a9-5499-a761...	s3://mast/level1/shots/120...	View More
11085531600000000000	MAST	Tokamak	12080	D	7bc9ac3e-3c72-5755-885d...	s3://mast/level1/shots/120...	View More

Demonstrated cataloguing and federation capacity

Proof of Concept: Phase 1 – Complete

FusionDLHomeUK Mast

UK Mast Data Browser

Environment:

Dev

Table:

MAST_DATA_CATALOGUE_LEVEL1_CORE_GOLD

Rows Loaded: 15969

CREATIONDATE	DEVICENAME	DEVICETYPE	SHOTNUMBER	MAINPLASMASPECIES	ID	URL	Action
1108494780000000000	MAST	Tokamak	12070	D	3bcdcb98-fe85-53e7-899b...	s3://mast/level1/shots/120...	<button>View More</button>
				D	be7ee686-bf92-5add-bd5...	s3://mast/level1/shots/120...	<button>View More</button>
				D	d49fb320-421c-5d6f-a7e4...	s3://mast/level1/shots/120...	<button>View More</button>
				D	5bf4430b-0b6d-5fec-9053...	s3://mast/level1/shots/120...	<button>View More</button>
				D	90c76767-2a57-56ed-b3ee...	s3://mast/level1/shots/120...	<button>View More</button>
				D	766191c4-00a5-5e34-af49...	s3://mast/level1/shots/212...	<button>View More</button>
				D	9d1675aa-ddad-5c61-a14...	s3://mast/level1/shots/120...	<button>View More</button>
				D	69ca76c1-6014-5b1d-a1c0...	s3://mast/level1/shots/120...	<button>View More</button>
				D	c996224d-bc41-576b-a6f7...	s3://mast/level1/shots/120...	<button>View More</button>
				D	f158001b-897a-5f81-8967...	s3://mast/level1/shots/120...	<button>View More</button>
				D	8eb70e4a-96a9-5499-a761...	s3://mast/level1/shots/120...	<button>View More</button>
				D	7bc9ac3e-3c72-5755-885d...	s3://mast/level1/shots/120...	<button>View More</button>

FusionDLHomeUK Mast

Row Details - MAST_DATA_CATALOGUE_LEVEL1_CORE_GOLD

Column	Value
CREATIONDATE	1108494780000000000
DEVICENAME	MAST
DEVICETYPE	Tokamak
SHOTNUMBER	12070
MAINPLASMASPECIES	D
ID	3bcdcb98-fe85-53e7-899b-b40822446e79
URL	<button>Download</button>
ENDPOINTURL	https://s3.echo.stfc.ac.uk
CONFINEMENTTYPE	MCF
SCIENTIFICMETHOD	Experiment
NEUTRONYIELD	
COMMENT	NO SUSTAINED PLASMA BREAKDOWN EVEN LATER THAN IN 12069. SCHEDULER DIED, DATA OK AFTER...
DIAGNOSTIC	

[← Back to Browse](#)

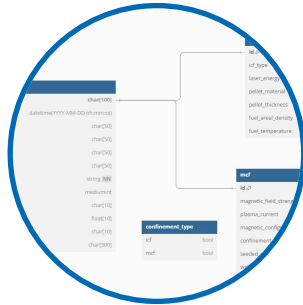
Proof of Concept: Phase 2 – In Progress

Goals: Expand the Fusion Data Lake with the LHD and Alcator C-Mod logs

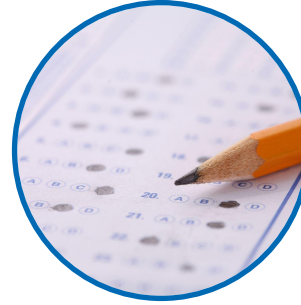
Timeline: September 2025 to January 2026



Networking and
Dataset ingestion



Data model mapping



E2E Testing

To demonstrate scalability!

Regulation: Terms of Service (ToS)



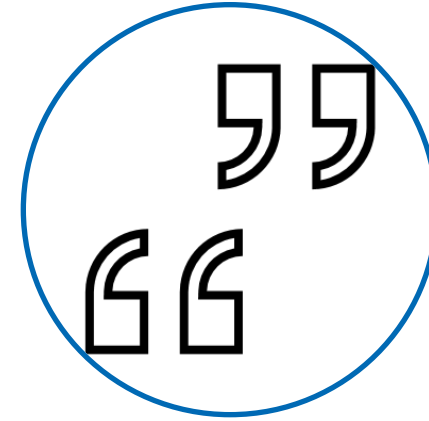
Uploads/Ingress

- Users who upload data must have the authority to share the data to the platform,
- Data must be licenced, either linked or in the metadata.



Downloads/Egress

- Data is only used and shared in line with ToS & data licensing,
- Data licenses will be transparent and accessible to enable compliance.



Attribution/Citation

- Data must be acknowledged in line with the data license,
- The platform must be referenced.

Regulation: Data Access Levels

Levels of data restriction:

1. Public – Open data. No login credentials required.
2. Internal – Open to anyone with login credentials to the platform.
3. Restricted – Only open to individuals with login credentials from institutions approved by the data owner.
4. Closed – Only be accessible to individuals approved by the data owner, with login credentials to the platform.

Requires a
[NUCLEUS \(IAEA\)](#)
account

Institutional
credential
authentication
using [OpenAthens](#)

You choose the level of access to your data!

Conclusions

The IAEA's Fusion Data Lake is being built to accelerate AI development in fusion research through Open Science and FAIR Data.

Progress:

- Scale able engineering design within standard fusion ontologies
- Proof of Concept – Phase 1 to Phase 2
- Developing regulation for an inclusive and safe platform



Come support the development and adoption of the Fusion Data Lake!

Thank you for listening!

Any Questions?

Daljeet Singh Gahle *et al*, IAEA
6th IAEA Technical Meeting on Fusion Data
Processing, Validation and Analysis
September 9th-12th, Fudan University, Shanghai

Conclusions

The IAEA's Fusion Data Lake is being built to accelerate AI development in fusion research through Open Science and FAIR Data.

Progress:

- Scale able engineering design within standard fusion ontologies
- Proof of Concept – Phase 1 to Phase 2
- Developing regulation for an inclusive and safe platform



Come support the development and adoption of the Fusion Data Lake!