

# EDITH Infrastructure

Deep Thinkers Meeting  
Rome, 16<sup>th</sup>-17<sup>th</sup> of May



EDITH is a coordination and support action funded by the Digital Europe program of the European Commission under grant agreement n° 101083771.



# A one-stop shop to discover, share, design, and use DHT

Our vision



EDITH is a coordination and support action funded by the Digital Europe  
program of the European Commission under grant agreement  
n° 101083771.



# Ecosystem

Establishing a common understanding of  
scientific catalogues and repositories



EDITH is a coordination and support action funded by the Digital Europe  
program of the European Commission under grant agreement  
n° 101083771.



# What to include?



**EDITH**

*Ecosystem for Digital Twins in Healthcare*

# A trinity of software

## Catalogue

- A place to share and discover research objects (publications, data, models,...)
- Unique (global) identifiers & versioning
- Metadata (manual, automated) to facilitate discovery
- Actual object may be available (stored) elsewhere
- Catalogue and harvesting services (federation, distribution)

## Repository

- Safely store and retrieve digital research objects (pubs, data, models,...)
- Files (versioning, metadata)
- Unique (global) identifiers & versioning
- Access policies (and sensitive data)
- Long-term preservation (storage classes)
- File & Sync (personal space)

## Platform

- Analyze, simulate, visualize, process, manage, interact, ...
- Software services (web apps, APIs, Jupyter notebooks, workflow engines, VDLs)
- Compute (HPC/HTC), storage and networking
- Collaboration and tiers
- Generic-purpose and domain-specific



# Main functionalities?

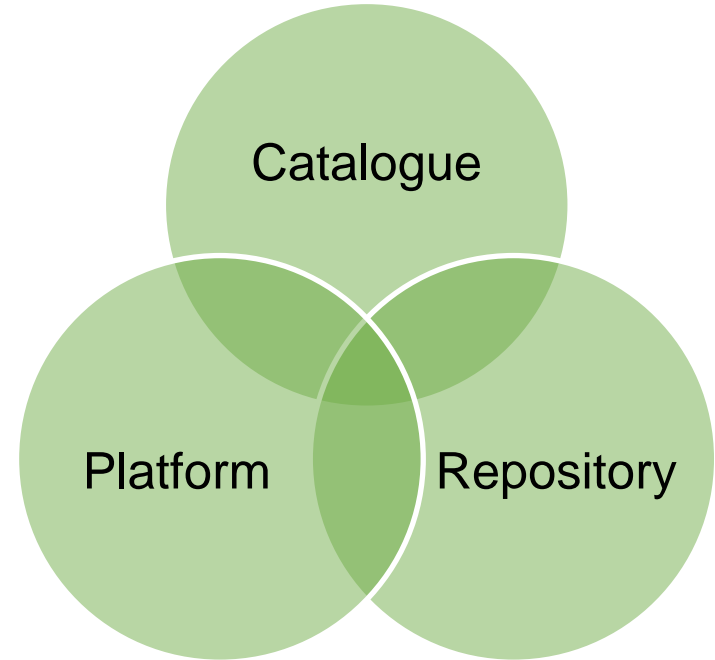


**EDITH**

*Ecosystem for Digital Twins in Healthcare*

# An ecosystem (and a future RI)

- Research objects in the Catalogue are stored to, and retrieved from, the Repository
- Research objects in the Catalogue are discoverable from the Platform and the Repository
- Research objects from the Repository and the Platform can be published in the Catalogue
- Research objects are linked between them in the Catalogue, Repository, and Platform
- Research objects in the Catalogue can be harvested by, and exposed to, other Catalogues



# A PoC of and springboard for ...

- Develop a distributed platform making available to users
  - 1) a federated repository of VHTs related resources,
  - 2) a combined set of open-source software toolkits, and
  - 3) access to computational services, enabling them to develop, test and integrate VHT models.
- Our vision and architecture is influenced by and informs this objective

**Platform for  
advanced Virtual  
Human Twin (VHT)  
models**

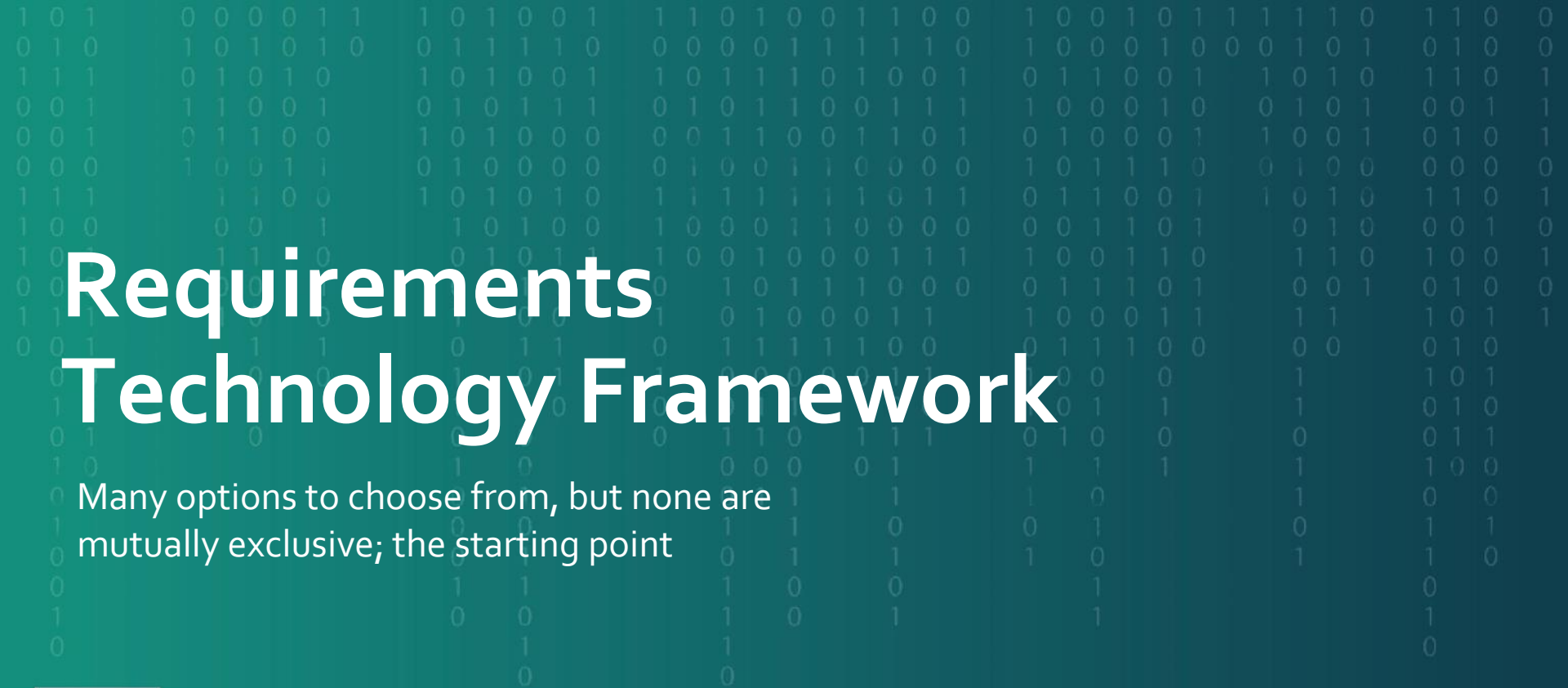
Digital Europe WP 2023-2024



**EDITH**

*Ecosystem for Digital Twins in Healthcare*





# Requirements Technology Framework

Many options to choose from, but none are mutually exclusive; the starting point



EDITH is a coordination and support action funded by the Digital Europe program of the European Commission under grant agreement n° 101083771.



# General directions



**EDITH**

*Ecosystem for Digital Twins in Healthcare*

# Cross-cutting directions (1/3)

- Open Source and Open \*
  - OS to be used exclusively for building, provisioning, operating the platform; any software to be provided as a service
  - Data and models can be proprietary (encouraged for industrial applications)
  - Commercial/Proprietary services (external) also supported (we are building an ecosystem)
  - Open Standards (support for proprietary standards only for interoperability)
  - FAIR guidelines; *'as open as possible, as closed as necessary'*



# Cross-cutting directions (2/3)

- Distribution and Federation
  - We need a common understanding on what these terms **mean** and what they **do not mean** at different levels and aspects of the VHT platform and lifecycle
- Distributed platform
  - Not centralized (all in one place), because not everything can be in one place (sensitive data) or are available (services); Assumption: a central node does exist
  - Trusted Research Environment (TRE)
    - Brings everything together, EU Health Data Space - compatible



# Cross-cutting directions (3/3)

- Federation
  - **We speak** the same language (metadata, API, standards, operating processes)
  - **We agree** to speak the same language, though **not all words** (i.e., not all services/capabilities need to be identical)
  - Different levels of federation can co-exist
- What needs to be federated
  - **Repository** (Data, Models) – **one** catalogue where they can be discovered
  - **Access** federated to build and visualize simulations
- What it means to be federated
  - Agreed (not necessarily common) Access Policies, SLAs and Monitoring
  - End-to-end observability, monitoring and accounting (how/where platform is used, is level of service OK, what resources shall we use, how to remunerate providers)
  - Heightened security (software toolkits deployed and hosted by each organization)
- We will deliver a PoC federation
  - Central node (EDITH stack)
  - EDITH partners (even for a few months)



# Catalogue



**EDITH**

*Ecosystem for Digital Twins in Healthcare*

# Catalogue (1/3)

- Research Objects (first-class citizens)
  - Data, models, notebooks, workflows, services, software, ...entire simulations (via Platform)
  - Single (with different views) catalogue or multiple catalogue instances (homogenized)
  - Specialized UI/actions/capabilities per RO (e.g., viewer, open with, add to my collection)
- Metadata schema
  - Base schema for all, with extensions for specific ROs (additional metadata)
  - Support for existing schemata (validation, store, ingest, transform, download base and original) – input needed
  - No need to reinvent the wheel or overly standardize right now (lower entry barrier, reuse/repurpose, expand)
  - Future support for integrations (e.g., EOSC)
- Identifiers
  - Catalogue-specific unique stable identifiers for ROs (and users)
  - Assign DOIs (automated for all or after curation)
  - Link with external user identifiers (e.g., ORCID)



# Catalogue (2/3)

- Files
  - Relevant for data and models (more?)
  - Deposited (browser, ftp, Repository, ...) or referenced (optional policy-based retrieval and archiving; sensitive data)
  - Open or closed vocabulary of file types (raw, archived)
  - Basic automated validation and sanitization (security); more services exposed from the Repository
  - Stored into the Repository (original and any derived versions); download original or automatically transformed (RO-specific)
- Additional 'resources'
  - Catch-all for anything that helps a user understand the RO (scope, purpose, application)
  - Linked or deposited (e.g., documentation, applications, other ROs in own or external catalogues)
- Publishing workflows
  - Self-served from registered users (authorized) following Helpdesk review/vetting (security, quality); user responsible for accuracy
  - Organization-level organization and publishing rights (e.g., uni->lab, delegated publishing process)
  - Tiered curation (metadata, data, models) and RO 'badge' (e.g., reference data, see next)
  - RO-specific publishing workflows (mix and match)





# Catalogue (3/3)

- License (linked with access policies)
- Access policy for metadata (visibility) and RO (download/use)
  - Public (for all), registered users only, authorized users only (groups, request), case-by-case (vetting), private
  - Embargo period, sensitive data, commercial/free, NDA
  - Metadata and RO handled differently; metadata generally open to allow machine-discoverability (EOSC, RIs,...)
- RO Badges & Tiers (not all are equal)
  - Reference data (expertly curated), Curation tiers (e.g., replicated), FAIR (e.g., automatically checked)
  - Crowd-sourced (e.g., ratings, comments)
- RO Derivatives (from another RO)
  - Depends on original license; original publisher or other; provenance and links



# Repository



**EDITH**

*Ecosystem for Digital Twins in Healthcare*

# Repository

- Multi-tier storage
  - File system, object storage, block storage
  - Federated and cloud-ready (S3)
  - Identifiers and versioning; backup and disaster recovery policies
  - Access policies (sensitive data, TRE, Platform)
- Enterprise File Sync and Sharing (EFSS)
  - Complete collaboration suite for file sharing (users, groups), web-accessible, multiple clients, large-file support
- Extra services
  - Schema validation (known schemas, externalize to Catalogue)
  - Profiling (automated and user-steered insights on data)
  - Ingestion
  - Anonymization and de-identification



# Platform



**EDITH**

*Ecosystem for Digital Twins in Healthcare*

# Platform

- Suite of loosely-coupled (and federated) software and services
  - Expandable by design; core building blocks
  - Process, analyze, visualize, simulate
  - Platform-wide access policies, tiers and quota
- Standardised Workflows
  - CWL Workflows authoring and execution (cloud, HPC) over multi-clouds; portable, replicable, and scalable
- Jupyter Notebooks
  - Interactive computing (cloud, HPC); portable and scalable
- VDI
  - Remote desktop (web-access); preconfigured and hardened (TRE for sensitive data)



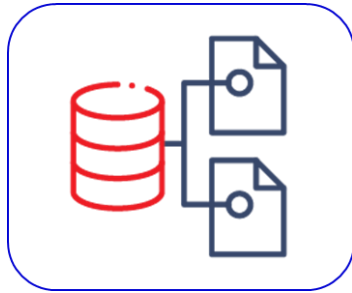
# User Profile and Access

- Idea: maintain distinction between user profiles and roles ( $m:n$  relationship)
- Several role categories defined:
  - patient/citizen
  - healthcare professional
  - developer
  - admin
- Authentication via existing identity providers
- Distinct GUIs adapted to each role
- Programmatic access to platform via API (preferably REST).

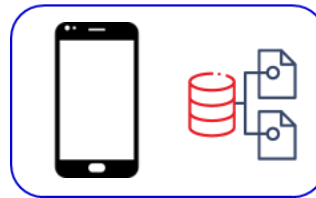


# Federation - paradigm shift

Centralised dataset



Multiple datasets



# Federated learning for Medical Informatics Platform (MIP) of EBRAINS

- Outline
  - Clinical data cannot be shared, transferred and stored in a centralized way
  - Data ownership remains fully within the boundaries of the contributor
- Approach
  - Hardened security and privacy standards
  - Dataset harmonization
  - Federated analytics





# Federated Learning - Medical Informatics Platform (MIP)

- Collaborative analysis of medical data while preserving privacy and security.
- Train models on decentralized data without centralizing sensitive information.
- Data privacy ensured: Encryption, differential privacy, and secure computation techniques.
- Secure infrastructure: Data owners can participate in training without sharing raw data.
- Standardized data representation: Compatibility and integration among institutions.



# Roadmap

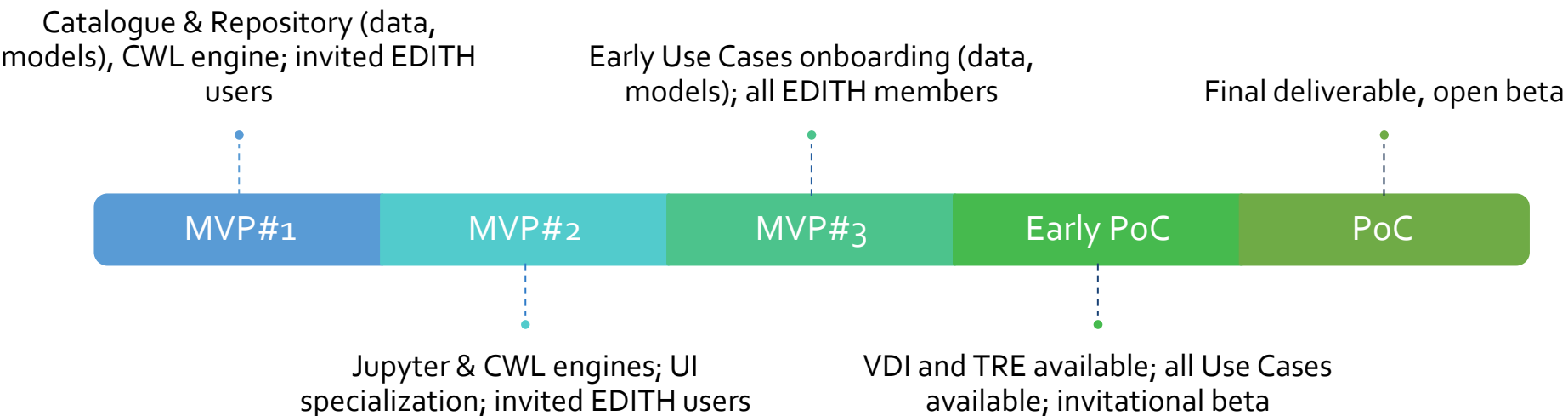
From requirements and design to  
the PoC and an MVP



EDITH is a coordination and support action funded by the Digital Europe  
program of the European Commission under grant agreement  
n° 101083771.



# Timeline



EDITH

Ecosystem for Digital Twins in Healthcare

# What do you think?



EDITH

EDITH is a coordination and support action funded by the Digital Europe program of the European Commission under grant agreement n° 101083771.

