

Break out session 2 – Tech Stack

Input for topic 2:

*Collection from multiple sources, integration across different systems,
interoperability*

EDITH CSA – Deep Thinkers meeting, Rome

May 16-17, 2023

Alfons Hoekstra, UvA

RESOURCES



models

data

compute

storage



models

data

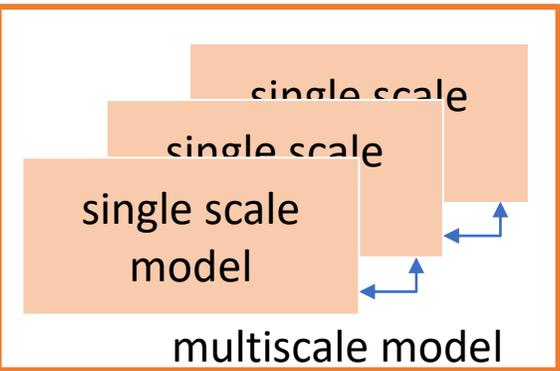
compute

storage

single scale model

⋮

single scale model



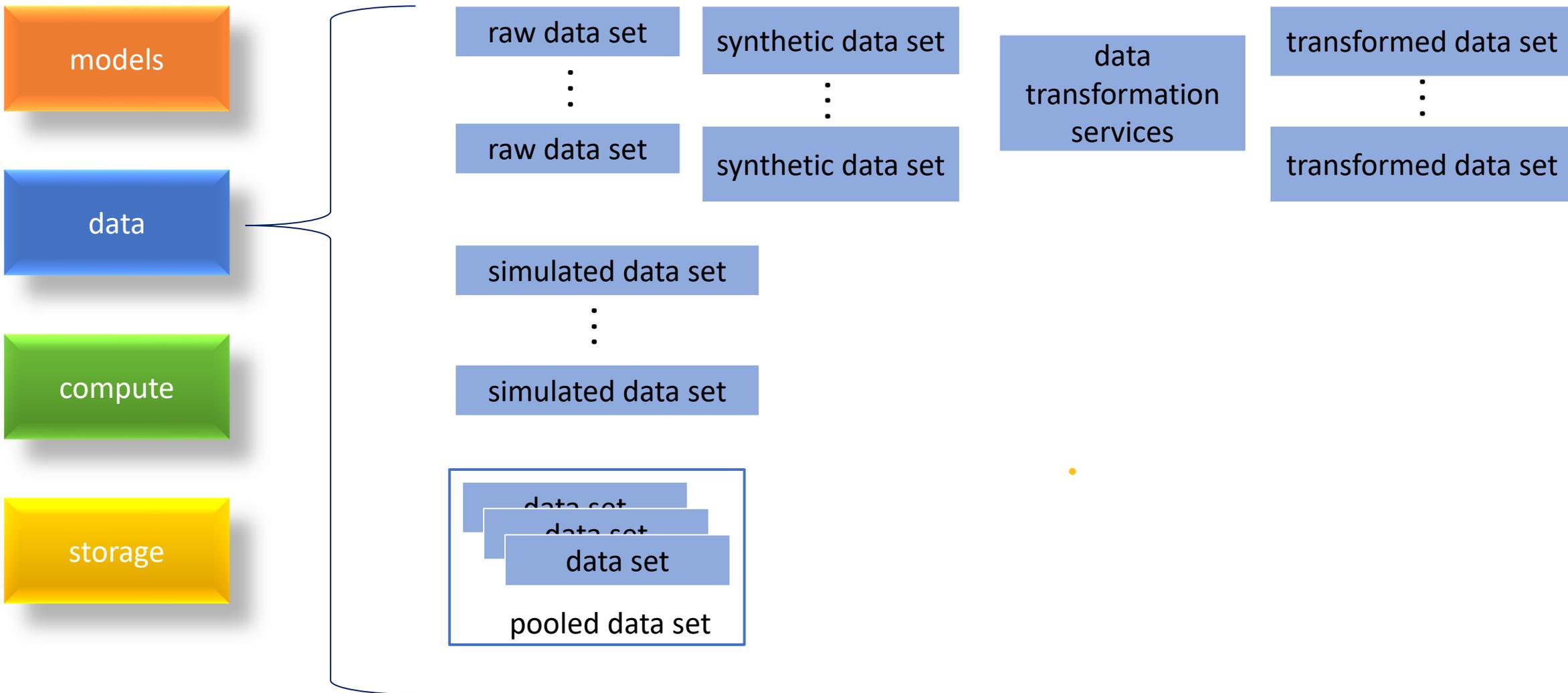
⋮

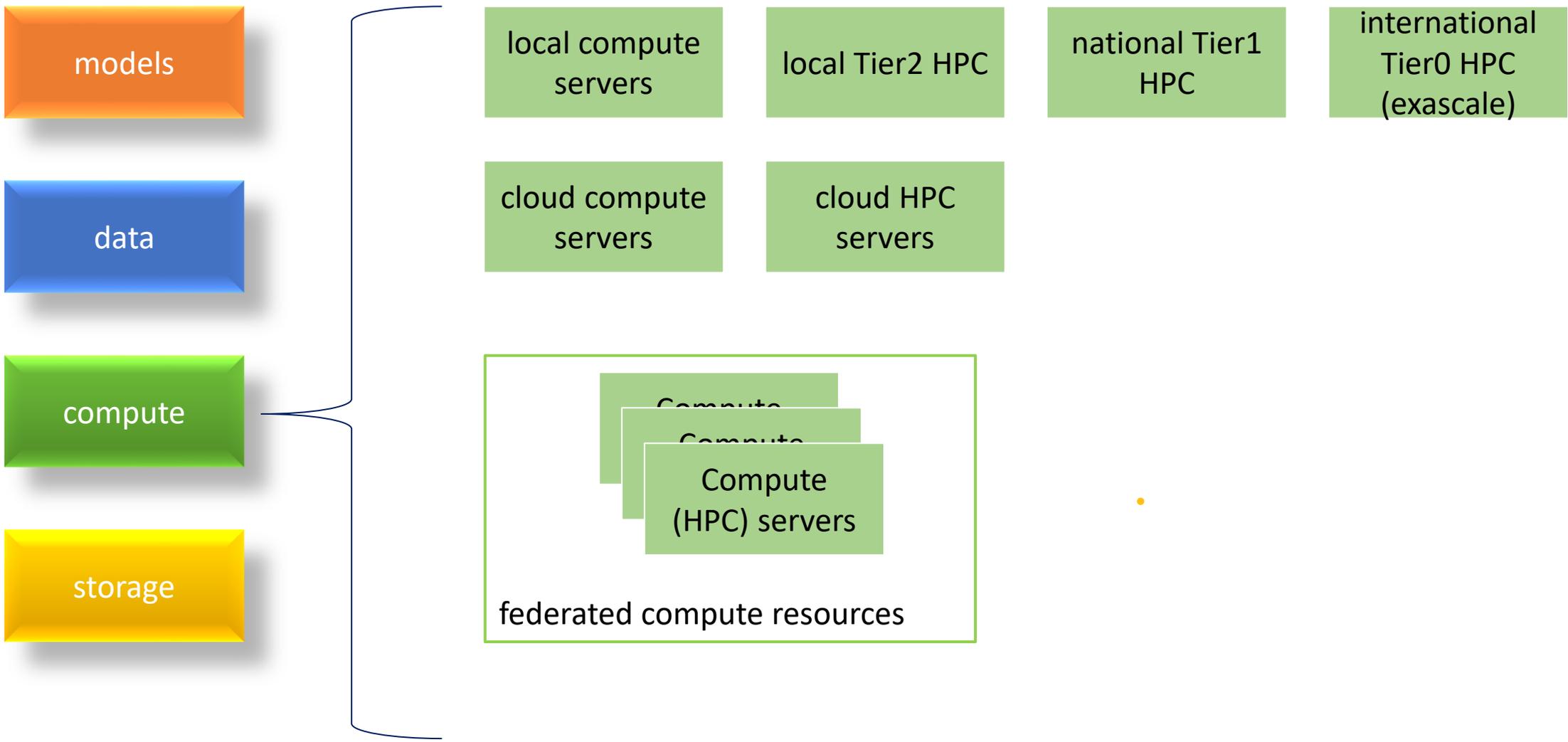
multiscale model

A model, in this context, is assumed to be a computational model that can be executed on a compute resource, so, an executable (stored somewhere).

multi organ
multi compartment
so, 'multi something'







models

data

compute

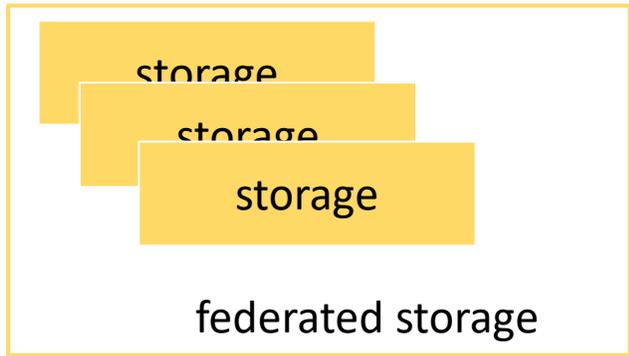
storage

local storage, not accessible from outside

local storage, accessible from outside

remote storage

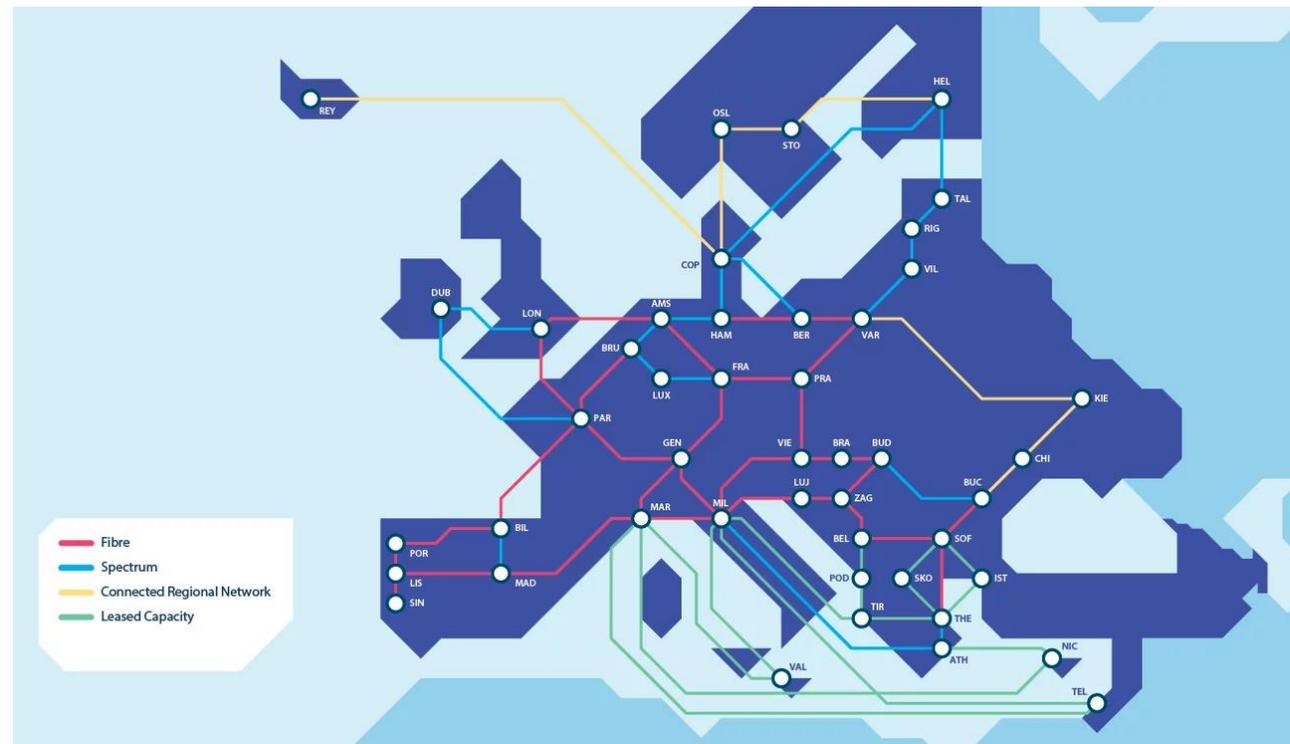
cloud storage



.

Let's not forget Networking

- Shifting very large volume's of data around may be needed and is not trivial.



Questions

1. Is this list of resources complete?
2. Naming conventions?
3. When do we call a resource a '*model*' and when do we call it a '*data transformation service*'?

Integration of resources on two levels

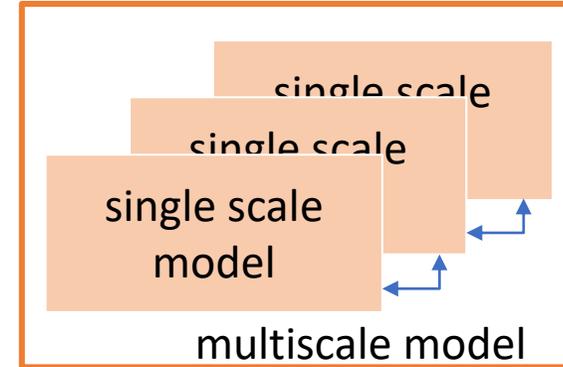
- *Inside* the models/data/compute/storage spaces
 - For *models*, e.g. integration of single scale models into multi-something models
 - For *data*, e.g. pooling of raw, synthetic, transformed, simulated data, including data transformation services, for (stratified) populations or individuals
 - For *compute* and *storage*, e.g. federating some local and remote resources
- *Between* models, data, compute, storage
 - This is actually needed to create a full blown DTH and to execute it.

Questions?

1. Are integrated resources again a single resource? Naming of those?

Model Integration

- Loosely coupled, one way
 - Workflows
- Tightly coupled, two way
 - More advanced methods are required
 - e.g. coupling libraries



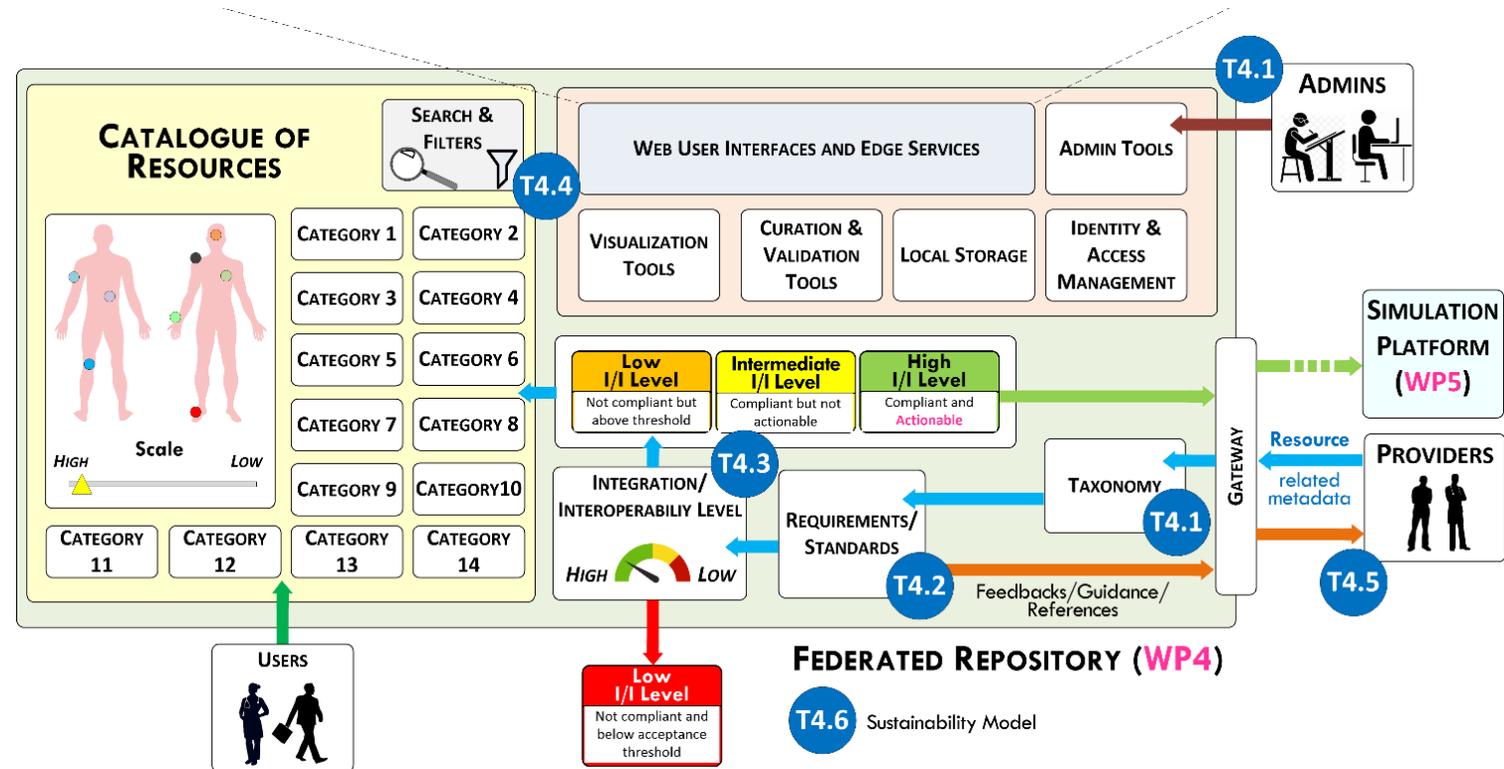
multi organ
multi compartment
etc

Data integration / pooling

- Standardisation of data input and integration
- Data interoperability of construction data
- Interfacing of distributed data.
- Metadata harmonisation and interoperability
- Integrating data from different sources (BRIDG, FHIR)
- sharing disease and phenotype information of a patient (e.g. Phenopackets)

Edith repository

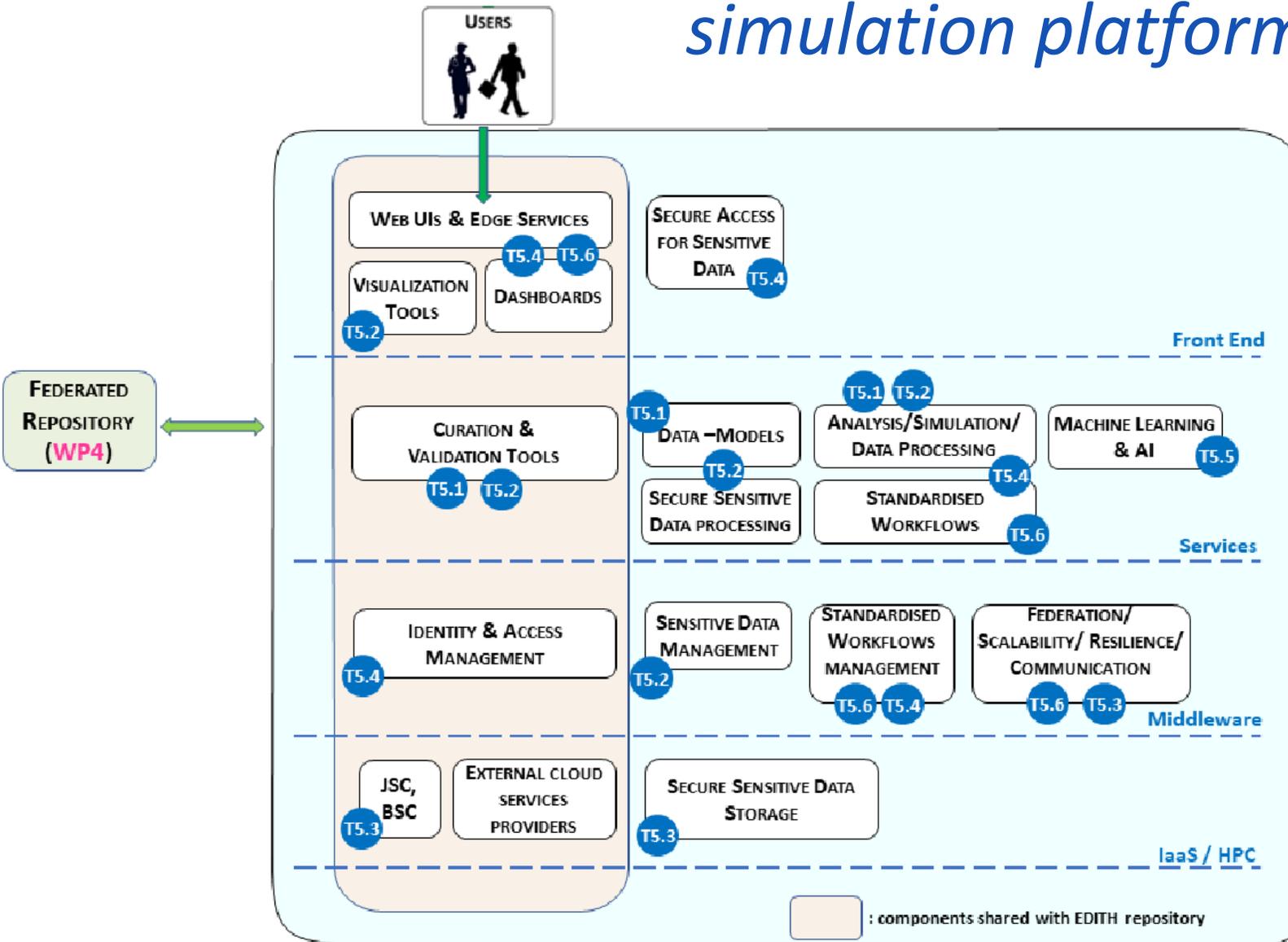
- Will be a virtual collaboration context composed of multiple users, sites, and organisations having the goal of pooling their VHT resources.



Compute / storage integration – federation - orchestration

- Lots of systems available, do we as community have specific demands that go beyond what is currently in production?
 - E.g. Fenix, EOSC, Gaia-X, OpenAir, HealthyCloud

simulation platform



VHT as a data crawler space

- DTH workflows automatically execute when (new) data becomes available
 - Eager data flow (as opposed to the reduction paradigm where a workflow would execute when it's output data is needed by another workflow).
1. What benefits would this bring?
 2. What are the technical implications?
 3. What could be technical solutions?

Break out session 2 – Tech Stack

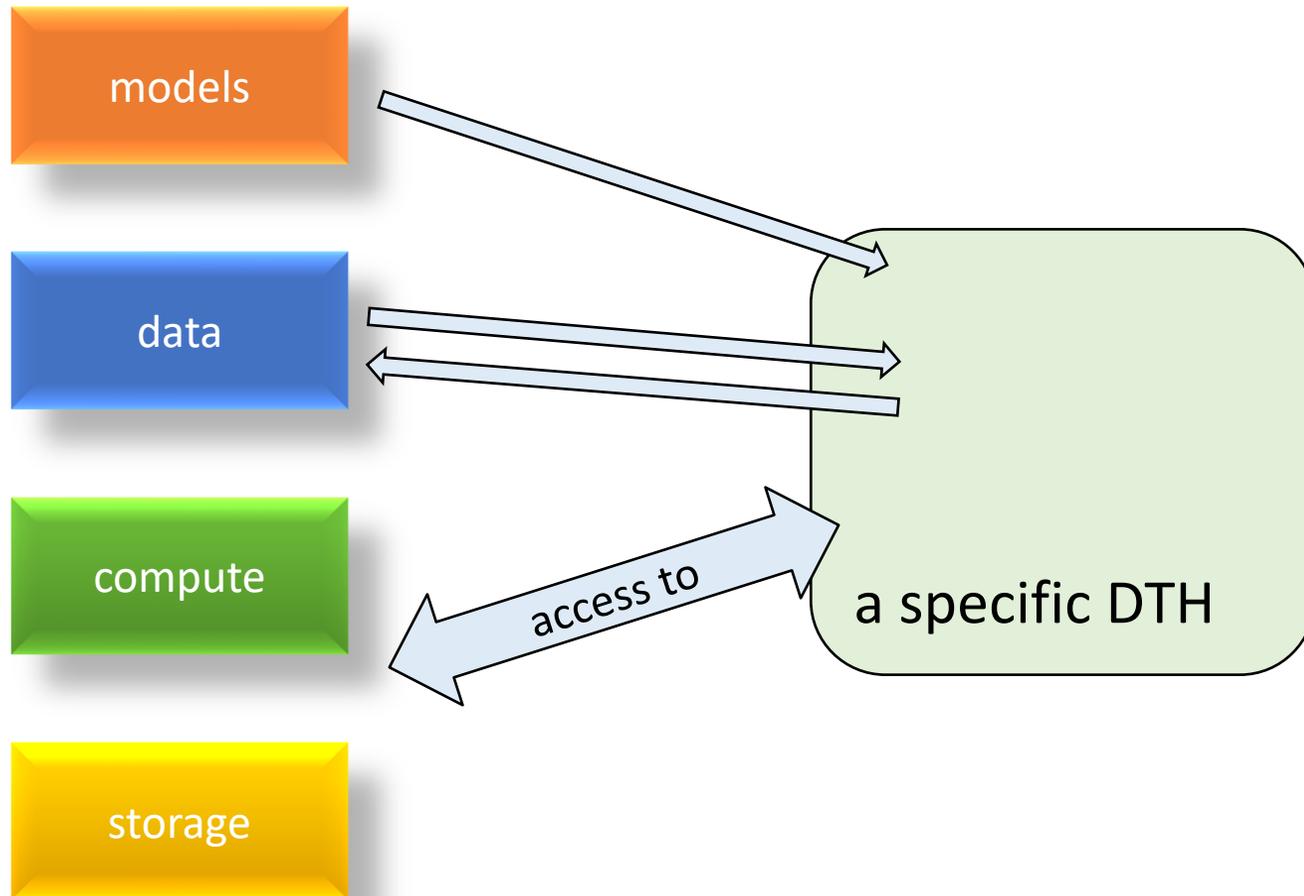
*Input for topic 3:
Workflows*

EDITH CSA – Deep Thinkers meeting, Rome

May 16-17, 2023

Alfons Hoekstra, UvA

Integration between models, data, compute, storage



Use workflows to achieve this.

•

Questions

1. Do you agree that we define a workflow as the combination of models and input / output data, dynamically *requesting* access to compute / storage / networking resources?

Scientific Workflows

- A scientific workflow system is a specialized form of a workflow management system designed specifically to *compose* and *execute* a series of computational and/or data manipulation steps, or workflow, in a scientific application.

Chee Sun Liew, Malcolm P. Atkinson, Michelle Galea, Tan Fong Ang, Paul Martin, and Jano I. Van Hemert. 2016. Scientific Workflows: Moving Across Paradigms. ACM Comput. Surv. 49, 4. <https://doi.org/10.1145/3012429>

- Lots of good systems out there, e.g.
 - From Jupyter notebooks to Kepler, to many others
 - The SANO/Cyfronet Model Execution Environment
 - Would be good to have a list
 - Will the VHT infrastructure support all (flexibility) or standardise on one (interoperability)?

Standardised workflow systems

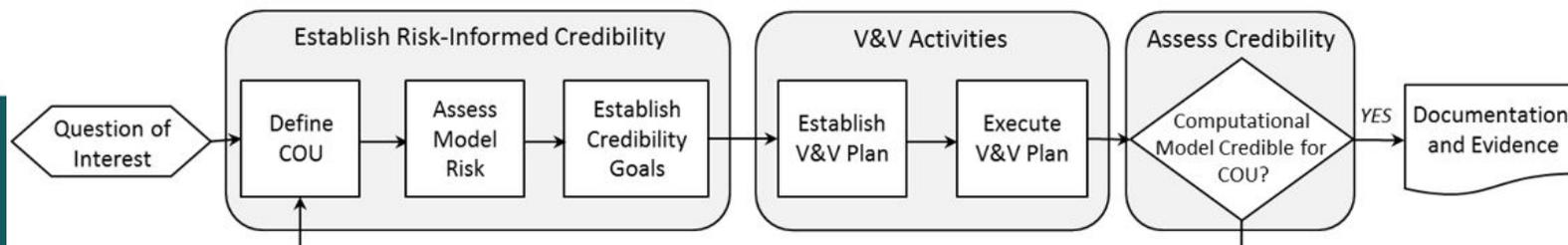
- Standardised Workflows and the Common Workflow Language (<https://www.commonwl.org>)
- Many tools already developed or implemented in other EC projects with involvement of many consortium partners
- Use-case driven collection of tools and components
 - Any results/conclusions that could go in the roadmap?
- List of existing SW to be used and integrated as is (vs. development of new components)
 - Is this list already available?
- Definition of FAIR workflows (CWL)

Questions

1. Will we support a single workflow system, e.g. CWL and build on that, or support requested system?
2. Should we strive for a VHT-workflow standard, leveraging existing standards?

Some more questions

1. Are there prototypical DTH workflows, or standard components for DTH workflows?
 - Generic DTH
 - Population specific DTH
 - E.g. having standard components that automatically check the CoU/QoI for which the DTH is validated and issue warnings when DTH is used outside that context?
 - Subject specific DTH
 - As above, and maybe other functionalities that kick in when moving from population to subject specific?
 - Or for UQ campaigns
 - E.g., maybe each DTH workflow could/should be equipped with automatic non-intrusive UQ (relying e.g. on easyVVUQ developed in the EU-funded VECMA project)?
 - Validation workflows according to V&V40?
 - following ASME workflow



Some more questions

1. Composing DTH workflows

- Using advanced user interfaces, manoeuvring atlases of human anatomy?
 - Of the quality of e.g. Elsevier's complete anatomy, <https://www.elsevier.com/solutions/complete-anatomy>
- Leveraging the 6D framework as backbone?
- Exploiting advanced knowledge graphs on human (patho)physiology?
 - E.g. Elsevier's Healthcare Knowledge Graph or Biology Knowledge Graph (see <https://www.elsevier.com/solutions/biology-knowledge-graph>) or comparable efforts.
- Re-using existing workflows, maybe even automating that?
- Advanced AI to help, e.g. a ChatGPT like interface, advanced search engines in all available data/models/literature, to propose templates of workflows to be further tailored by DTH developers?

Some more questions

1. Executing DTH workflows

- Completely automated, hiding all complexity from the DTH user?
- Automatically sending jobs to most suitable compute resources, pulling data from the right locations, moving data around, invoking dedicated networking infra, etc?
- For HPC jobs, advanced reservation, collocation of pooled resources, etc?
- What are VHT specific demands in this respect, if any? Security of data? Other?