Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

National DataLake Integration

Diego Ciangottini - on behalf of Spoke3_WP4-5

INFN
Istituto Nazionale di Fisica Nucleare

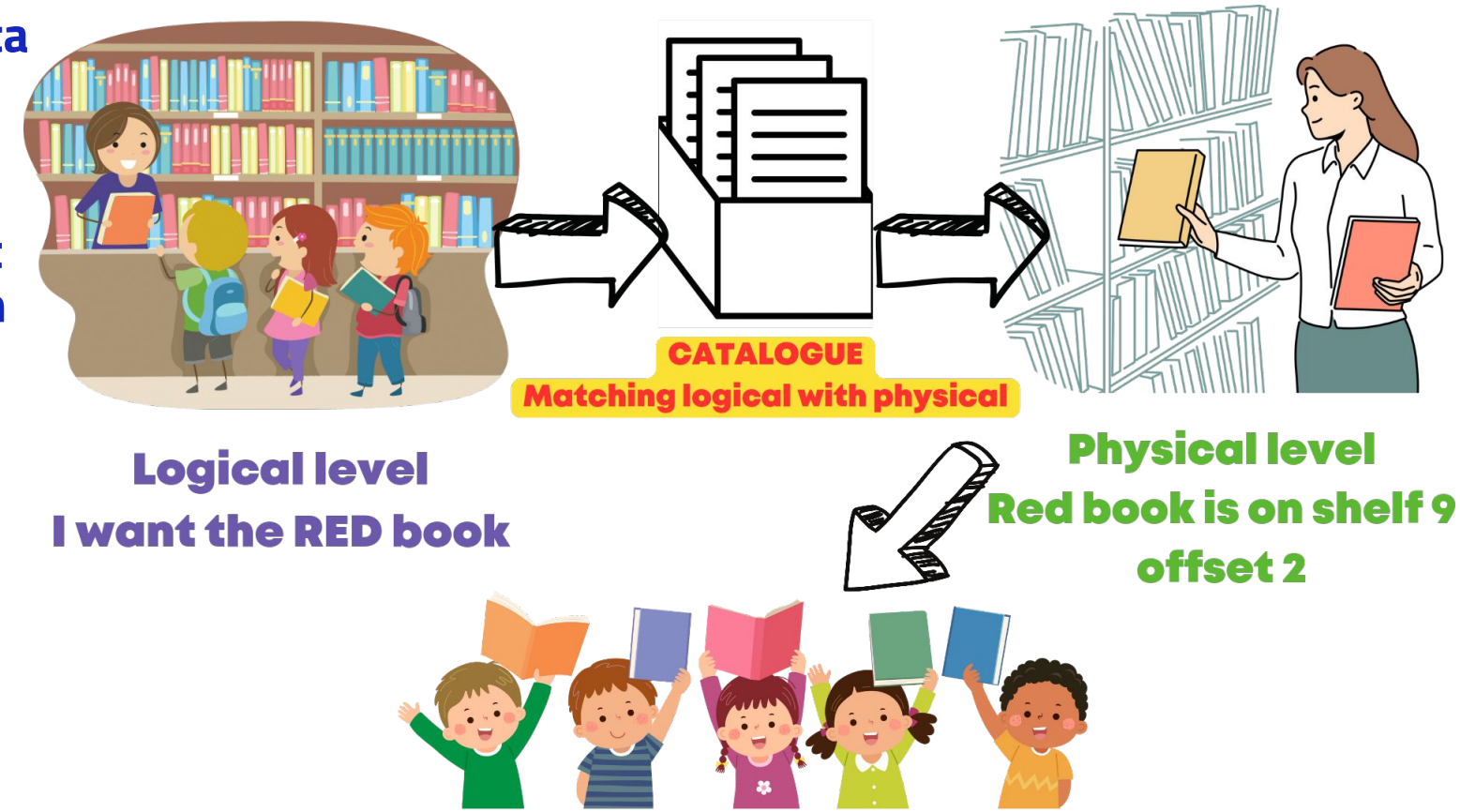
Spoke 3 Technical Workshop, Trieste October 9 / 11, 2023

What do we want to achieve

A data lake for science use cases

- Declarative management of the data replication and lifecycle
- Get it to work for different science use cases
- Separate the physical management of data from their logical utilisation

Ideally the science stakeholders should be able to leverage the same data management system with changes only at the catalogue level interface.

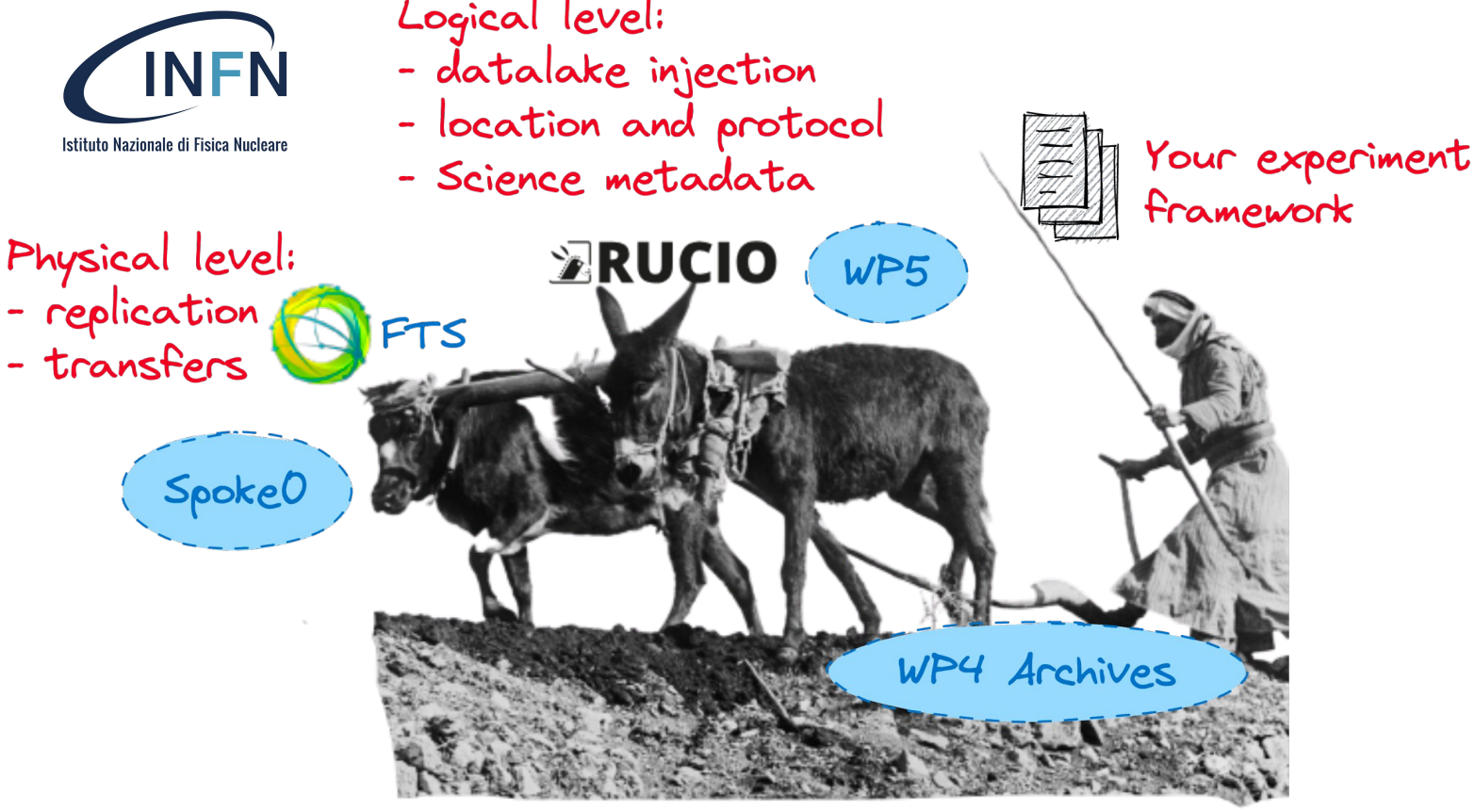


In other words…

Start from INFN DataCloud experience and evolve it based on user requirements

Being particularly careful on responsibility sharing e.g.:

- **we might expect Spoke0 to manage data transfers and storage site (+ Spoke3 services at regime)**
- **WP5 for integration of RUCIO development services and archive databases**
- **WP4 for archives plugin integration on the frameworks**



The plan

Physical level:

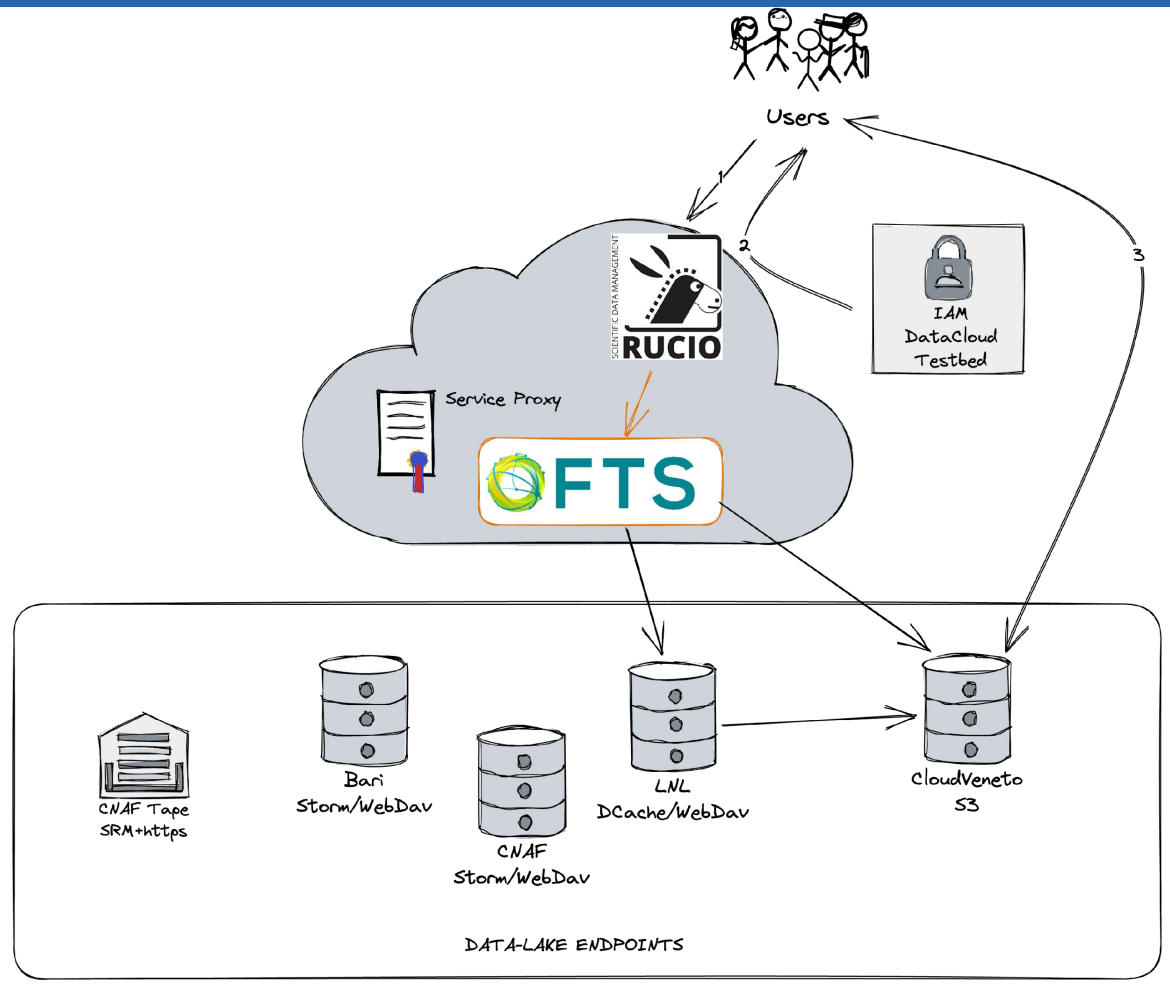
- Get initial space for a playground on a couple of storage sites
- Protocols matter: S3/object + an HTTP/WebDav might be a good first set

Physical+Logical matching (WP5 <-> INFN DataCloud):

- Setting up dev services required for managing transfers and data accessibility (FTS+RUCIO)

Interface level:

- Make Rucio interfaces with metadata databases used by archives



N.B. we acknowledge the existence of strong synergies with Spoke2 and Data-cloud initiatives, we should work in strict coordination to avoid waste of efforts.

Current status

See next presentation… anyhow, in a nutshell:

- **We started to look at the details on how the logical levels can work together with "external" databases for metadata**
- **Just started a training process where we can leverage the INFN RUCIO testbed to get hands dirty with the data management RUCIO experience**

We need to understand better our use cases to plan for a proper integration

Next stops

- Have a first volunteer use case to provide a working code example
- Provide the software for allowing the first prototype of integration b/w RUCIO and archive metadata
- Planning for “scale out” phase, with more use cases