Green Paper

Reference Architecture of the EU FCT Trusted Research Data Ecosystem

March 2025

This report is a summary of research conducted in the context of the LAGO project and reported in Deliverable 3.3, titled 'Reference Model and Architecture of the EU FCT Trusted Research Data Ecosystem'.

The LAGO project, Lessen Data Access and Governance Obstacles, aims to 'address the data issue in the FCT research landscape by building an evidence-based and validated multi-actor reference architecture for a trusted EU FCT Research Data Ecosystem'. The project aims to lay the foundations for a trusted European FCT RDE.

The LAGO consortium comprises 25 partners across 14 countries. LAGO is funded by the European Union under grant agreement number 101073951.

LAGO: Lessen Data Access and Governance Obstacles. Funded by the European Union. Grant agreement number 101073951. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Table of Contents

Foreword

LAGO provides a trustworthy and sustainable Research Data Ecosystem (RDE) to address the lack of domain-specific data of sufficient quality and quantity, which is essential for the appropriate training and testing of tools developed within the FCT research landscape.

LAGO outlines the fundamental building blocks for establishing a trustworthy and effective EU Research Data Ecosystem within a tightly regulated framework that adheres to ethical and legal requirements.

To this end, the Reference Model and Architecture of the LAGO RDE have been developed to specify:

- Procedures for creating high-quality, reusable datasets for FCT research.
- Protocols for managing and governing data, ensuring access and sharing.
- Mechanisms for data exchange between different organisations within the RDE.
- Standards and practices for modelling and publishing metadata to facilitate efficient data discovery and usage.
- Procedures to ensure compliance with legal and ethical standards, addressing security and privacy concerns, and safeguarding data integrity and confidentiality while meeting relevant regulations and standards.

We acknowledge the support of the EU-funded LAGO project in enabling this work.

Ernesto La Mattina
LAGO Project Coordinator
Head of Data Centric AI Research Unit—AI & Data Lab
Engineering Ingegneria Informatica SpA

RDE: Reference Model and Architecture

The LAGO Research Data Ecosystem (RDE) provides a unique environment that guarantees access to FCT research data for researchers, Law Enforcement Agencies (LEAs), practitioners, and any other security stakeholders. This access is facilitated through standardised procedures and controls that ensure security, trust between participants, high data quality, and compliance with legal, ethical and regulatory requirements.

Compared to other data-sharing initiatives, the value of the LAGO ecosystem lies in its holistic approach. It not only covers the standards, procedural, and technical aspects of data sharing (as is often the case in other data spaces or ecosystem projects) but also incorporates mechanisms that ensure trust. Participants are accredited within the ecosystem by a trusted authority, ensuring that each participant can be confident that they are engaging with other trusted entities. Additionally, LAGO supports data governance even before the data-sharing process begins, providing guidelines for compliant data creation, annotation, and quality assessment, as well as mitigating the risks associated with sharing data that might raise security concerns.

Main benefits include:

- **Data preservation and availability**: data within the ecosystem is catalogued and remains accessible to participants, even after the end of a research project, preventing valuable datasets from being lost or rendered invisible.
- **Data sovereignty**: the ecosystem ensures that
  - Datasets are stored on the provider's premises and are only shared when requested.
  - Identities and related credentials are securely stored by the owners and shared only when necessary to interact with other participants.
- **Standardised data sharing**: the ecosystem provides standardised procedures and protocols for documenting and sharing data in a fully controlled manner, ensuring compliance with legal requirements through contracts between data-sharing parties.
- **Guaranteed security and trust**: only accredited organisations are allowed to participate in the RDE. Accreditation involves the **trusted authority** issuing credentials to organisations after evaluating their trustworthiness. These credentials serve as proof of trust and can be presented to other participants in the ecosystem.
- **Technical interoperability**: LAGO offers a set of standards based on EU-adopted and open standards, ensuring compliance with the *openness* principle. These standards describe information about datasets, credentials, and participants, and standardised APIs allow for seamless communication between technical components within the ecosystem so that anyone can develop its own implementation.
- **Transparency**: the adoption of an Ethereum-based ledger ensures transparency by logging all activities within the ecosystem, such as the publication of metadata, data sharing between participants, and contract agreements.

The **Reference Model** [1] and **Architecture** [2] of the LAGO **Research Data Ecosystem** (RDE) consist of the following core elements (Figure 1), which interact through standardised protocols and procedures:

- **Participants**: *data providers* and *consumers*.
- **Resources**: research datasets and tools for co-creation and assessment.
- **Procedures**: guidelines for effective participation in the RDE.
- **Standards and protocols**: for interoperable resource exchange.
- **Technical components**: enabling participants to access research resources.

1. The *LAGO Reference Model* is a high-level conceptualisation of the proposed solution, focusing on the main elements and actors, how they interact with each other, and in which processes they participate. It provides a common representation to allow stakeholders and technicians to easily understand the context through a simple description of the model and the desired goals explained using unambiguous terms.
2. The *LAGO Reference Architecture* derives from a Reference Model, translating its concepts into concrete building blocks for a sustainable solution. Thus, the functionalities of the Reference Model are mapped into (eco)system decompositions, procedures, guidelines, and best practices.

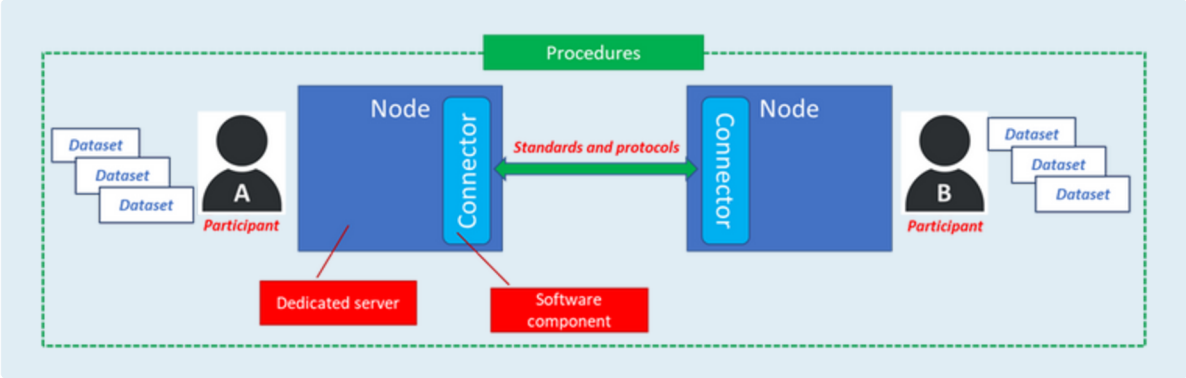

Figure 1: The LAGO RDE Core Elements

The implemented reference architecture is based on key design principles, including *decentralisation*, *data sovereignty*, *data quality*, *openness, transparency, trust, interoperability*, and *portability*. These principles ensure that research data is made available by federated entities while maintaining full control with data providers over which data to share, to whom, and under what conditions (such as usage licences and policies). High-quality datasets are essential for FCT research as they are needed to train and test data-driven solutions. The approach ensures openness through clear rules, specifications, and protocols for data sharing and transparency regarding data handling. Trust is built through confidence in the identity and capabilities of participants, while interoperability and portability are achieved by enabling data exchange via technical means and standardised protocols. This is done while ensuring compliance with *ethics*, *legal* standards, *privacy requirements*, *proportionality*, and *risk assessments*.

The proposed architecture adopts a **hybrid decentralised-centralised data repository**, which allows data providers to retain full control over their data. Providers decide what data to make available, to whom, and under which terms (i.e., licences and usage policies).

Participants in the LAGO federated ecosystem can store their datasets on their dedicated servers, referred to as ***Nodes***, and provide access via specific software components known as ***Connectors***.

To make datasets discoverable, data providers register metadata about their datasets in the ***Catalogue*** (Figure 2). This metadata includes not only the nature of the datasets but also the usage policies and licences under which they are made available.

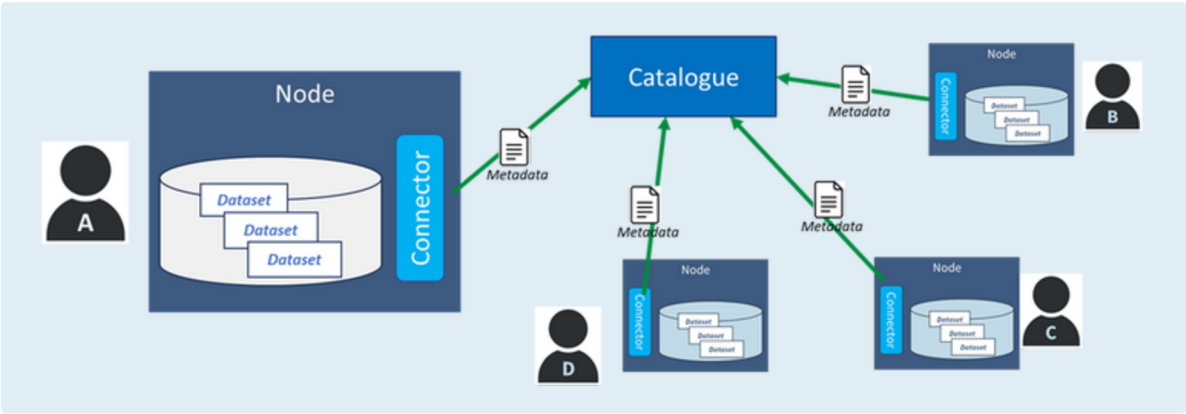

Figure 2: Metadata Published on the Catalogue

Participants can *query* the ***Catalogue*** to find datasets of interest. To request access to a specific dataset, a participant must send a request to the data provider (Figure 3). If a *licence* has already been specified, the data consumer must accept the licence terms before requesting access. In cases where no licence is specified, the data provider may require the establishment of a *contract* with the data consumer before sharing the dataset. This contract will outline the purpose of the data sharing, the terms and conditions for data exchange, usage policies, and any necessary clauses to protect both parties and prevent data misuse.

Once the licence terms are accepted or the contract is established, the dataset can be shared.

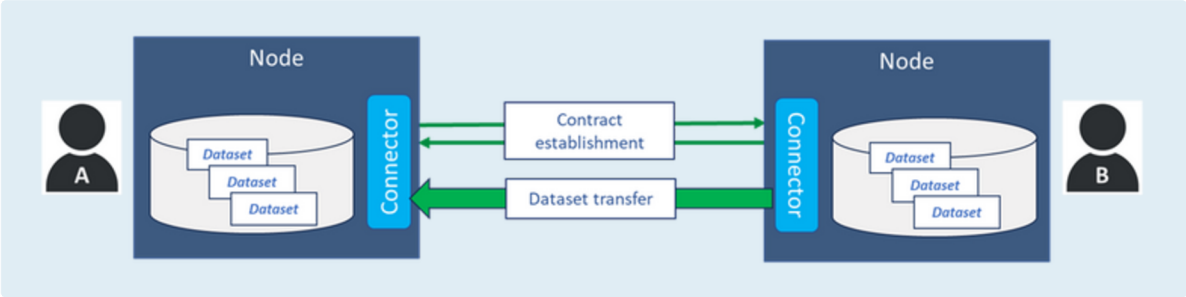

Figure 3: Contract Establishment and Dataset Transfer

Usage policies can be enforced either through *technical* means (e.g., duration, number of downloads) or controlled at the *organisational* level. The ***Connectors*** are responsible for monitoring the validity of those usage policies that can be *technically* enforced, ensuring that the data sharing process respects the agreed contractual terms.

To ensure full transparency, every transaction within the RDE is logged on the ***Ledger*** (Figure 4). The ***Ledger*** records events within the ecosystem, such as new participant registrations, metadata entries in the ***Catalogue***, contract stages, data sharing between participants, and instances of non-compliance with contracts. Given its critical role in tracking events across the RDE, the ***Ledger*** is distributed across federated nodes and is based on blockchain technology, specifically Ethereum-based technologies.

The ***Ledger*** comprises *Ethereum full node*s, which maintain the complete blockchain, and *Ethereum light nodes*, which rely on full nodes for chain information. To ensure accurate tracking of all activities within the ecosystem, each RDE component is paired with an Ethereum node, with its own Ethereum account registered.

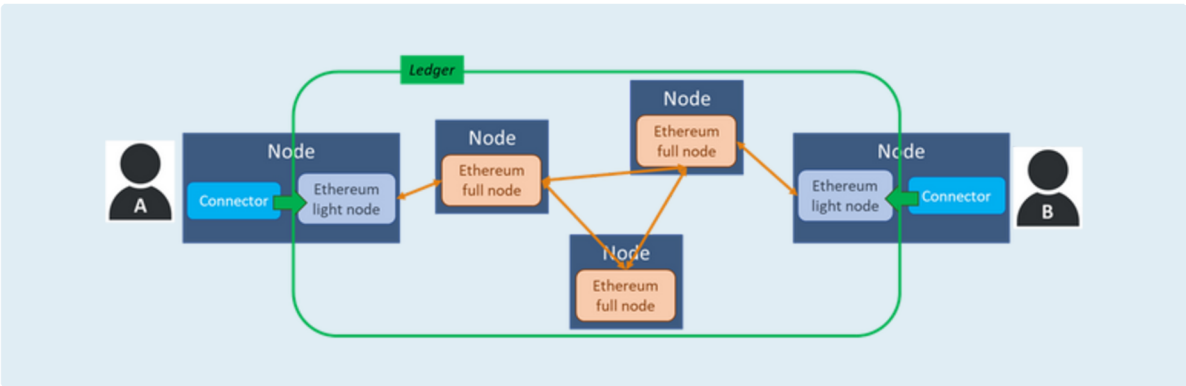

Figure 4: Ethereum-Based Ledger in the RDE

Participant Onboarding

To build trust among the various stakeholders involved in research and to address concerns about the potential risks of data sharing, an organisation can only become a participant in the Research Data Ecosystem (RDE) after undergoing an Onboarding Procedure, carried out by a trusted authority, known as the *Issuer*.

When a new onboarding request is received, the *Issuer* performs the necessary checks to verify the claims made by the participant. These checks involve an offline assessment of the trustworthiness of the requesting organisation, such as reviewing its previous work, ensuring there are no legal infringements, verifying certifications, and evaluating the purpose of the organisation's participation in the RDE. If these checks are successful, the request is approved.

The onboarding procedure results in the verified participant receiving:

- *Credentials* that confirm their trustworthiness.
- A *certificate* that can be used for signing contracts within the RDE.

Additionally, the *participant's information* is published in the ***Catalogue***, making it searchable by other participants.

In line with the *interoperability* principle, RDE credentials adhere to the Verifiable Credentials Data Model [3]. The ***Issuer*** sets up a dedicated ***Node***, which hosts an Ethereum node to log accreditation activities on the ***Ledger***. The ***Issuer*** also sets up a custom ***Connector*** (the ***Issuer Connector***) to manage accreditation requests and the generation and signing of credentials (Figure 5).

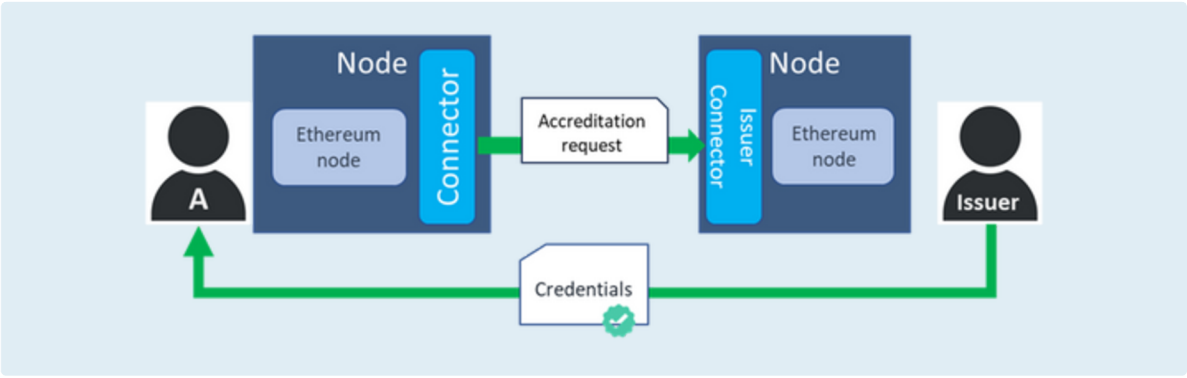

Figure 5: Accreditation of a New Participant

3. Sporny, M., Thibodeau Jr, T., Herman, I., Cohen, G., Jones, M.B., 2025. *Verifiable Credentials Data Model v2.0*. W3C Proposed Recommendation, 20 March. Available at: https://www.w3.org/TR/vc-data-model-2.0/ [Accessed 30 March 2025].

The credential status is recorded by the ***Issuer*** on the ***Ledger*** and can be used by anyone to verify the participant's credentials. Credentials can be registered with a VALID, REVOKED, or DEPRECATED status. The ***Issuer*** can update the status when necessary. For example, if a participant withdraws from the RDE, its credential status can be set as REVOKED. Similarly, the status can be DEPRECATED if certain claims (such as expired certifications) are no longer valid.

Before interacting with another ***Participant***, their ***Connector*** will request the participant's credentials (Figure 6). The credentials are then validated by verifying the signatures of both the ***Participant*** and the ***Issuer***, as well as checking the credential validity status on the ***Ledger*** by querying the Ethereum nodes. This ensures that participants can trust that the other party has been accredited by the ***Issuer***. After successful verification, the two ***Connectors*** authenticate each other and exchange an authentication token, which will be included in all future requests between them.

Figure 6: Credential Verification for Mutual Authentication

A similar process applies when a ***Connector*** interacts with the ***Catalogue***, whose credentials are provided during the setup of the ecosystem. In this case, the ***Participant Connector***:

1. Verifies the validity of the credentials.
2. Authenticates with the ***Catalogue*** using the verified credentials.
3. Sends a request to the ***Catalogue*** to register the participant's information.

Once the ***Participant's*** information is registered on the ***Catalogue***, the ***Participant Connector*** is ready to manage and exchange datasets within the RDE.

Metadata Cataloguing and Search

In accordance with the *Interoperability* principle, LAGO has developed a standard vocabulary called LAGO-DCAT-AP [4] for metadata representation. This is an extension of DCAT-AP.

Among other classes, the most relevant for LAGO are:

- ***Resource* class**: a resource described by a metadata record in the catalogue.
- ***Dataset* class**: an extension of *Resource*, including additional metadata to describe the dataset.
- ***Distribution* class**: represents an accessible form of the dataset, such as a downloadable file.

According to **DCAT-AP**, a *Dataset* can have one or more associated *Distributions*, representing the different formats in which the dataset is released. For example, a dataset containing temporal records could be available in formats such as **CSV**, **PDF**, **TXT**, etc.

LAGO-DCAT-AP includes additional metadata:

- Metadata about the dataset's *provenance* (sources and methodologies for collection, generation, annotation, anonymisation, transformation).
- Metadata describes *societal, legal, ethical, privacy, storage,* and *security aspects*.
- Data *sensitivity* (e.g., personal, anonymised, or pseudonymised data) and associated *biases, risks,* and *limitations*.
- *Purpose, intended use, known uses*, and *restrictions* on usage.

The publication process involves participants automatically extracting initial metadata from the uploaded dataset (e.g., size, format), refining the metadata (e.g., through a checklist or wizard), performing a **data quality assessment** (to ensure high-quality datasets are shared for research), and conducting a **risk assessment** (to prevent the disclosure of risky data). Participants also include a licence that specifies the terms and conditions for dataset usage and control visibility of the metadata for specific participants.

Data quality and *risk assessments* can be carried out independently of the publication process using standalone tools. ***Participants*** complete these assessments before making a dataset available in the RDE.

Data providers may make their datasets available through a ***licence***, meaning that any data consumer must accept the terms and conditions specified in the ***licence***, or through a ***contract***, meaning that both the provider and consumer need to agree on customised terms and conditions, signed by both parties. The data provider may decide whether or not to include the licence in the metadata to be published on the ***Catalogue***. If no licence is included, a contract will need to be established whenever a consumer requests access to the dataset.

Once the metadata is completed, the ***Participant*** can publish it on the ***Catalogue*** through their ***Connector*** *[7]*. The ***Participant Connector***:

1. Mutually authenticates with the ***Catalogue*** using its own credentials.
2. Sends the dataset metadata to the ***Catalogue***, including the authentication token generated from the mutual authentication process.
3. For each distribution, registers the metadata about the offered distribution on the ***Ledger***, such as owner, hash value, and licence, for notary and non-repudiation purposes.

The ***Participant Connector*** also allows participants to delist published metadata from the ***Catalogue*** at any time (e.g., when a participant wants to update the metadata of their datasets or permanently retire a dataset from the RDE).

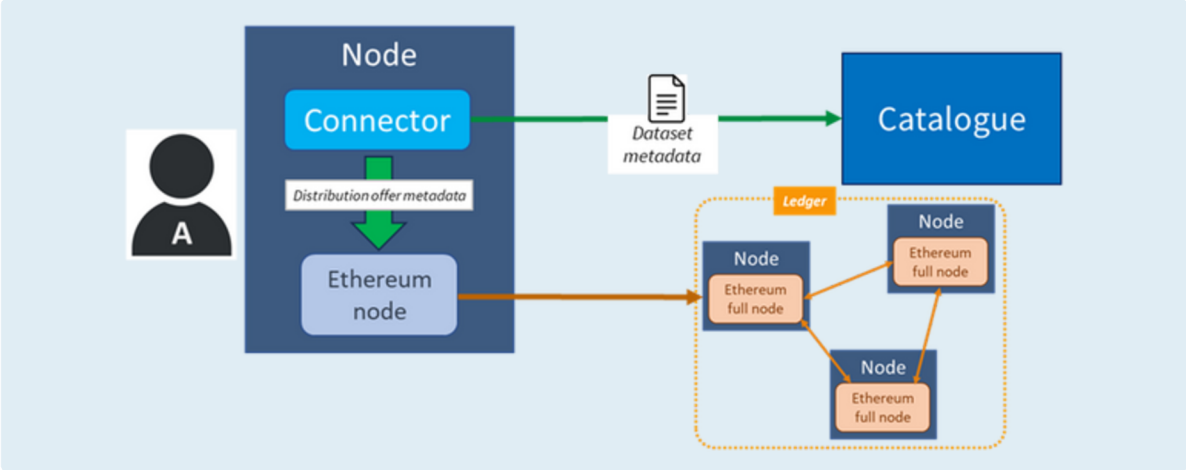

Figure 7: Metadata Publication on Catalogue and Ledger

Afterward, the ***Catalogue*** stores the received metadata. Another ***Participant*** needing datasets can search and view metadata of datasets registered on the ***Catalogue*** through the ***Participant Connector***. In this case, the enforcement of mutual authentication between the ***Participant Connector*** and the ***Catalogue*** ensures security and trust. The ***Participant*** can specify search criteria (e.g., keywords or metadata filters) and submit the search request through the ***Connector***, which is responsible for interacting with the ***Catalogue***'s APIs to retrieve the requested information.

Request, Agreement, and Sharing

The dataset request process involves several steps, ultimately resulting in an agreement between the participants exchanging the dataset.

The agreement can take two forms:

- A consumer accepts the licence if the provider has specified one in the metadata of the dataset.
- A contract is established between the provider and the consumer if no licence is specified by the data provider.

Interactions for data requests, agreements, and transfers always occur between two ***Participant Connectors*** *[8]*. According to established procedures, the two ***Connectors*** must first mutually authenticate before any further interactions can take place.
The request process then continues as follows:

- The consumer sends a request to access a specific distribution from their ***Connector*** to the provider's ***Connector***. The request must include the reason the consumer needs access to the distribution. If a licence is specified in the metadata, the request must also include the consumer's acceptance of the licence terms.
- The provider must then evaluate the access request. At this stage, it is strongly recommended that the provider perform another risk assessment, considering the requesting consumer's details (e.g., type of institution, certifications held, intended use, etc.).
- If the consumer has already accepted the licence, the provider's ***Connector*** logs the provider's acceptance on the ***Ledger*** and notifies the consumer's ***Connector*** that the request has been approved. If no licence was specified, a contract must be agreed upon between the provider and the consumer:

    a. The provider prepares and uploads a contract offer (in PDF format) to their ***Connector***.
    b. The provider's ***Connector*** automatically electronically signs the contract using the provider's private key, attaches the provider's certificate (issued during onboarding), and sends it to the consumer's ***Connector***.
    c. If the consumer successfully verifies the electronic signature, they can proceed to evaluate and accept the terms of the contract. The consumer's ***Connector*** then electronically signs the contract, attaches the consumer's certificate, and sends it back to the provider's ***Connector***. Meanwhile, the consumer's ***Connector*** logs the acceptance of the contract on the Ledger.
    d. If both electronic signatures are successfully verified by the provider's ***Connector***, the provider concludes the contract. The provider's ***Connector*** logs the conclusion on the ***Ledger*** and notifies the consumer's ***Connector*** that the contract has been successfully concluded.

- The consumer can now download the requested distribution:

    a. The consumer's ***Connector*** verifies the existence of the agreement on the ***Ledger***.
    b. The consumer's ***Connector*** logs the data access event on the ***Ledger***.
    c. The consumer's **Connector** sends the download request to the provider's ***Connector***.
    d. The provider's ***Connector*** verifies the existence of the agreement on the ***Ledger***.
    e. The provider's ***Connector*** logs the data exchange event on the ***Ledger***.
    f. The provider's ***Connector*** sends the requested distribution to the consumer's ***Connector*** via a secure encrypted channel (e.g., HTTPS).

From this point, the consumer has a copy of the requested distribution and can use it in accordance with the terms of the agreed licence or contract. In the case of a contract, both participants will have a copy of the double-signed contract stored on their premises at the end of the process.

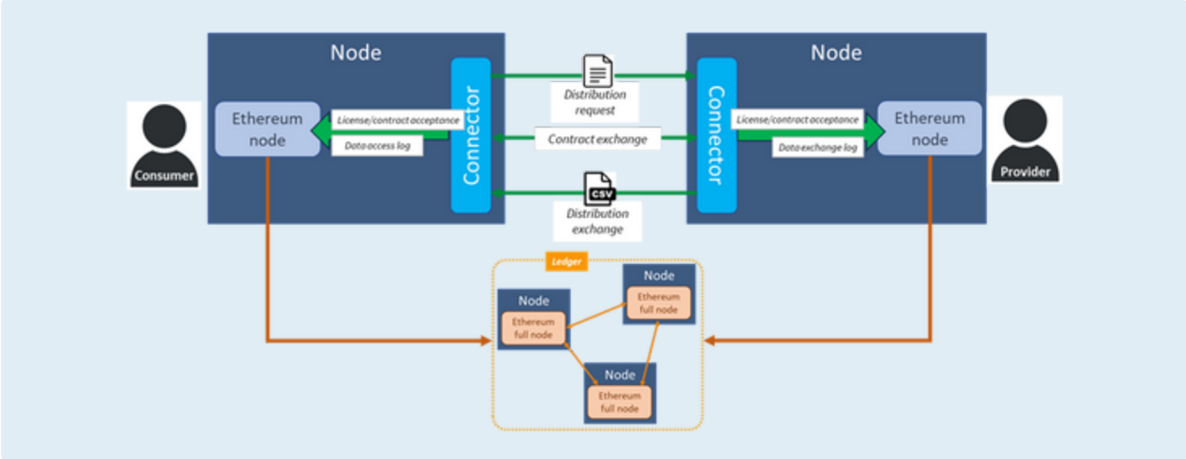

Figure 8: RDE Components Interaction for Data Exchange

Sharing Datasets

LAGO includes a *Centralised Storage* to facilitate the provision of datasets produced by short-lived entities (e.g., research projects) or organisations wishing to make their datasets available before leaving the RDE.

The ***Centralised Storage*** is managed by a ***Centralised Storage Administrator***, and organisations can request credentials from the ***Administrator*** to access it (Figure 9). Organisations do not need to be RDE participants to access the ***Centralised Storage***; they only need to be accredited by the ***Administrator***.

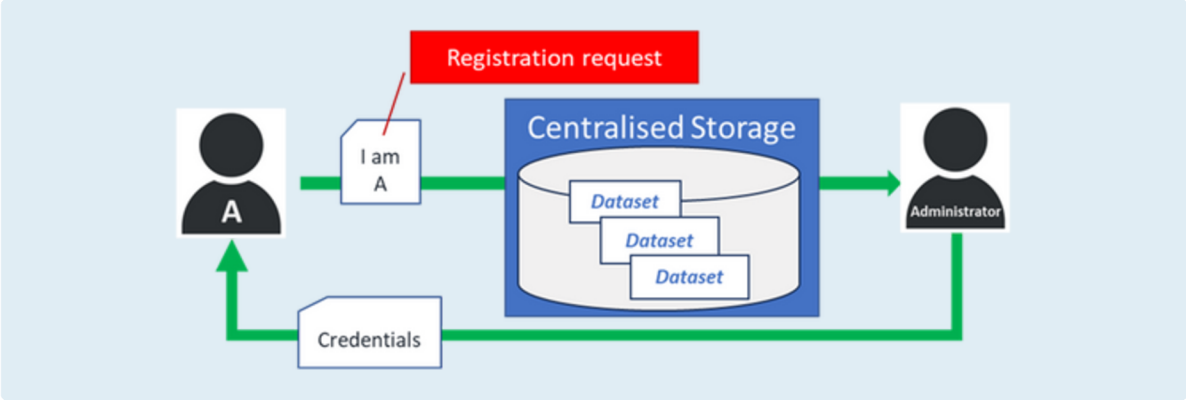

Figure 9: Organisation Requesting Credentials for the Centralised Storage

Once registered, organisations can view, upload, edit, or delete only their own datasets and publish metadata on the ***Catalogue*** (Figure 10). Datasets hosted on the ***Centralised Storage*** are provided with a licence, so no contract is required between provider and consumer. Once metadata is published, RDE participants can access these datasets after accepting the associated licences.

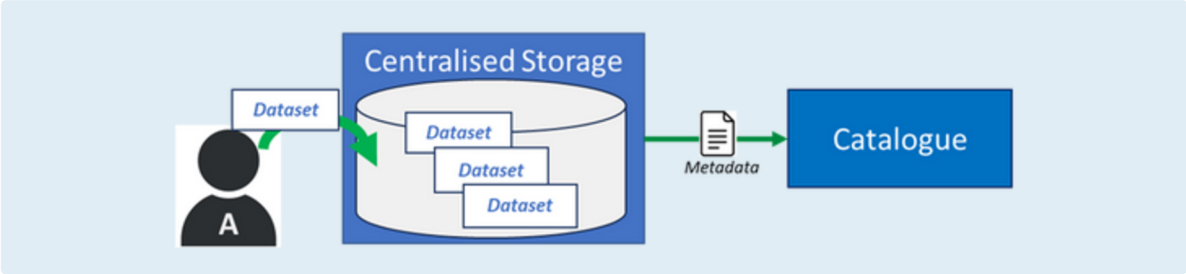

Figure 10: Dataset Upload on the Centralised Storage and Metadata Publication on the Catalogue

The Sandbox Environment

Finally, LAGO includes a *sandbox environment* designed to raise awareness of and encourage the use of research models made available within the ecosystem. This isolated environment is controlled and maintained by a *Sandbox Maintainer*.

Any RDE participant can send a request to the ***Maintainer*** to deploy a research model in the sandbox. Requests are sent from the ***Participant Connector*** to the ***Sandbox Connector***, a customised connector accessible by the Sandbox Maintainer and deployed on the RDE sandbox (Figure 11).

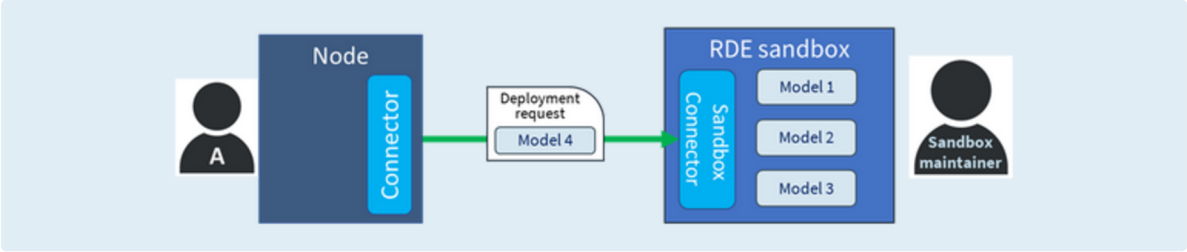

Figure 11: Sending Model Deployment request to Sandbox Maintainer

The ***Sandbox Maintainer*** is responsible for evaluating the proposed model's compliance with RDE principles before deploying it in the sandbox, as well as ensuring that the model does not pose any cybersecurity risks. Once approved, the research model is deployed in the sandbox, and metadata on the ***Catalogue*** is updated accordingly (Figure 12).

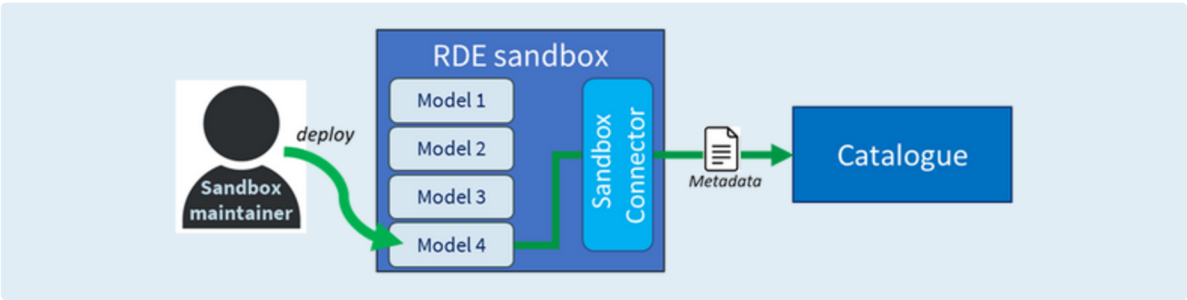

Figure 12: Model Deployed in the Sandbox and Metadata Published on the Catalogue

Any RDE participant can search for models in the ***Catalogue*** and test them in the sandbox. If interested, the participant can request a copy of the research model from the provider to deploy on their premises, following the same procedure used for requesting a dataset. Once exchanged, the participant can deploy the model as it is or into a ***private sandbox*** on their premises (Figure 13).

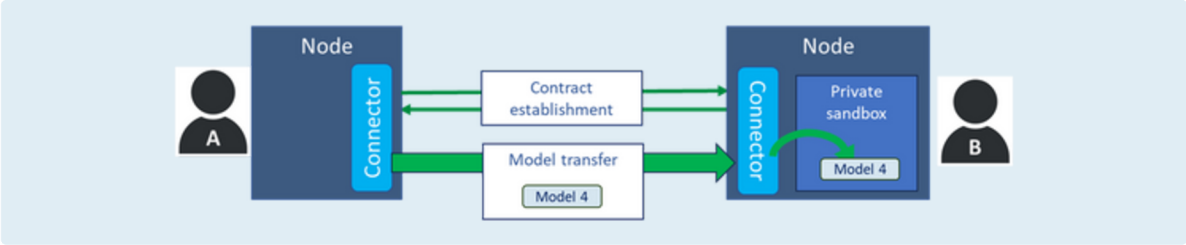

Figure 13: Model Exchange

The overall picture of the envisioned Reference Model is reported in Figure 14.

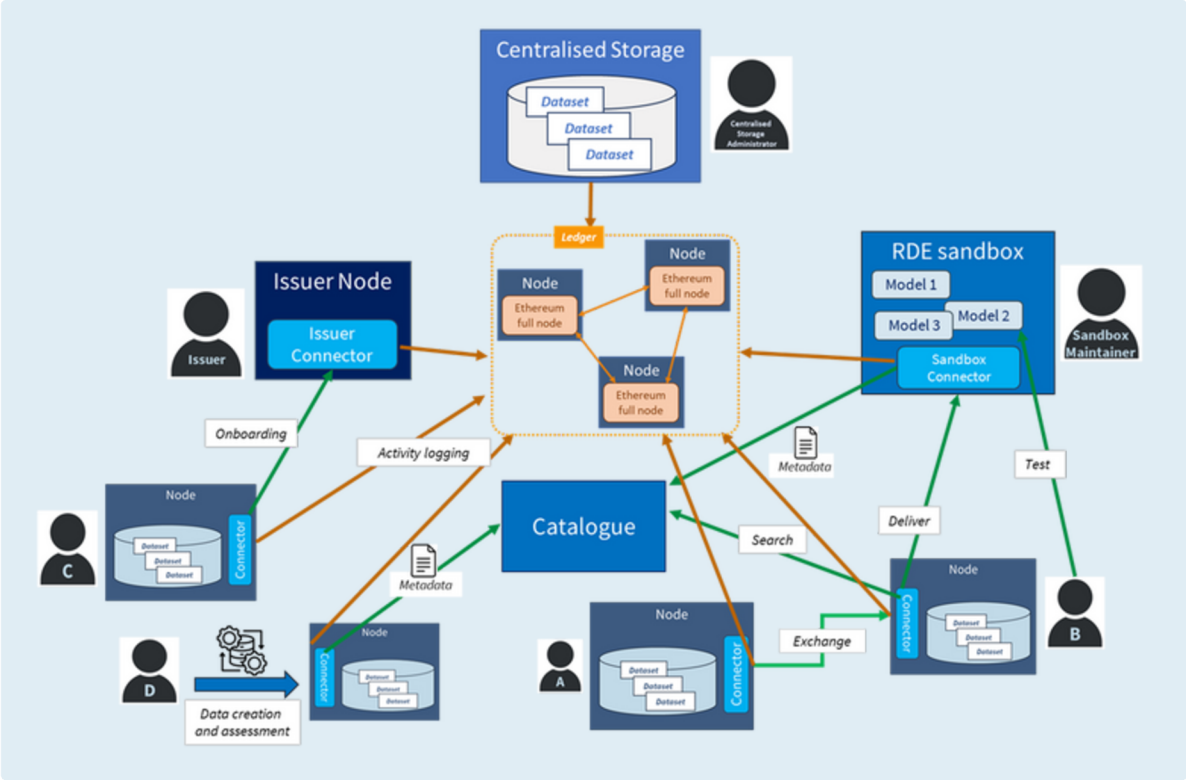

Figure 14: Research Data Ecosystem –Reference Model

Conclusions and Next Steps

The LAGO Reference Model describes various dimensions of the Research Data Ecosystem, including governance, legal, and technical aspects, along with recommendations for future work.

The **Reference Architecture** of the LAGO RDE, described in this green paper, represents the foundation for the future operationalisation of the RDE, which will require further effort before becoming fully operational. The LAGO project is currently defining a dedicated roadmap for this process.

The **Reference Architecture** provides the necessary guidelines, standards, and protocols to make the RDE function effectively.

A **Reference Implementation** of the LAGO RDE, which translates the **Reference Architecture** into a concrete and fully functional ecosystem, has been delivered and tested both internally and through cross-project use cases. The ecosystem is designed to be easily customisable and extensible to meet the needs of the entire security research community, addressing areas such as **Resilient Infrastructure**, **Border Management**, **Cybersecurity**, **Disaster-Resilient Societies**, **Strengthened Security Research**, and **Innovation in the Fight Against Crime and Terrorism**.

The adopted architectural approach allows for the further evolution of the RDE to address emerging needs and technological innovations over time.

Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
DCAT-AP	Data Catalog Vocabulary Application Profile
EU	European Union
FCT	Fight against Crime and Terrorism
RDE	Research Data Ecosystem
VC	Verifiable Credential
LLM	Large Language Model

eng.it/en