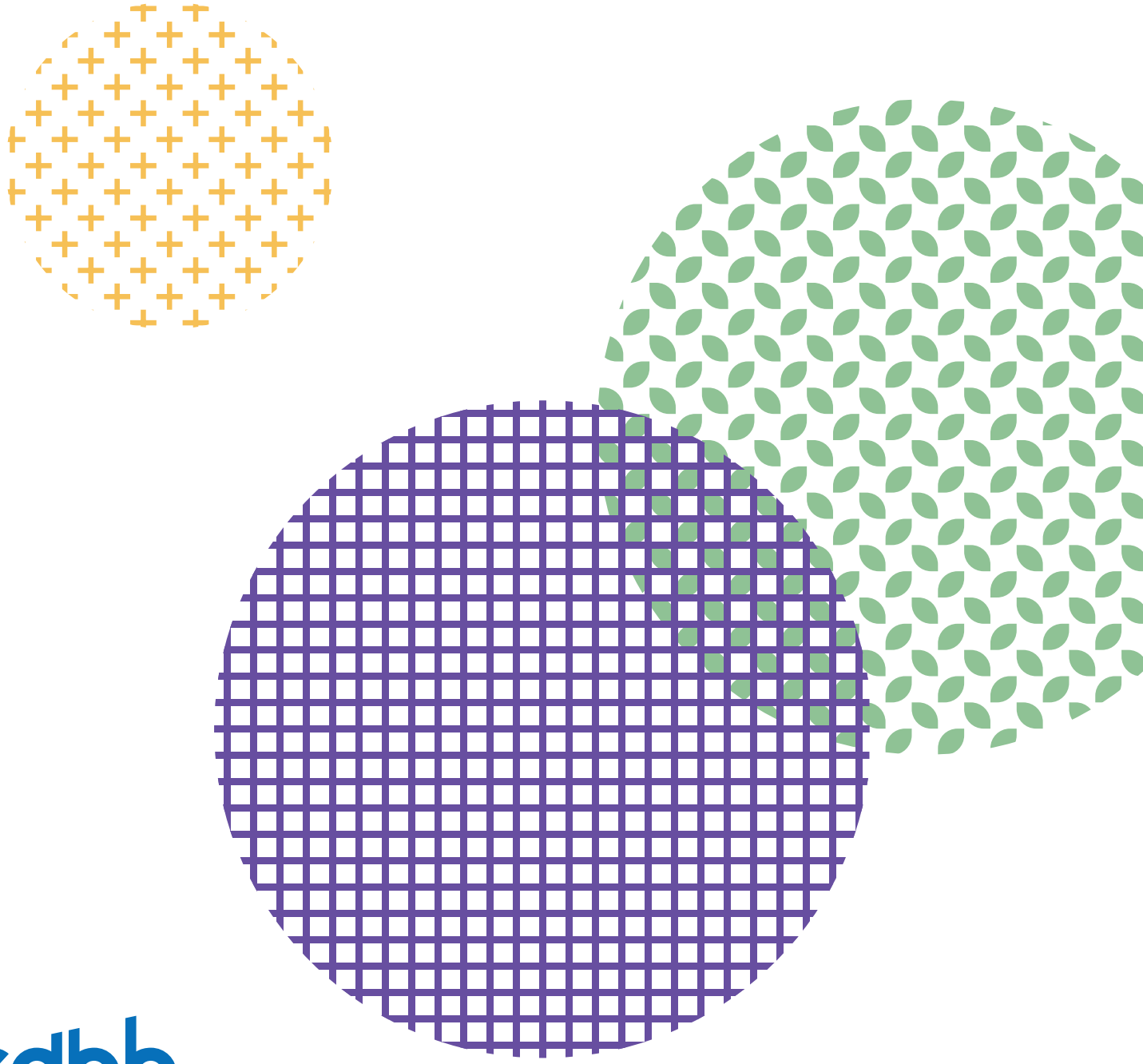
NATIONAL DIGITAL TWIN PROGRAMME

National Digital Twin

Integration Architecture Pattern and Principles



cdbb
Centre for Digital Built Britain

Contents

Executive summary

This document builds on the National Digital Twin (NDT) Programme's The Pathway Towards an Information Management Framework: A Commons for a Digital Built Britain (Hetherington and Matthews 2020), hereafter called the Pathway report. Enabled by the Construction Innovation Hub, it specifies a set of principles and a re-deployable architectural pattern composed of functional components for the creation of an Integration Architecture. This pattern, alongside the Reference Data Libraries (RDL) and the Foundation Data Model (FDM), will realise the National Digital Twin (NDT) Information Management Framework (IMF) vision. The re-deployable pattern allows the publication, protection, discovery, query and retrieval of data that conforms to defined Reference Data Libraries and the Foundation Data Model.

Three general architectural pattern options are explored (centralised, distributed, and federated), all of which could feasibly be used to realise the NDT vision. The benefits and concerns for each pattern are discussed with respect to the requirements. The recommended architectural pattern is a hybrid of these three approaches – centralising certain functions, whilst distributing and federating others. The main components that are required to provide the functionality of the Integration Architecture and their interactions are described.

The recommended architectural pattern will allow datasets to be shared securely so that they can be accessed by organisations that have a legitimate interest and contract in place. The publishers and consumers of datasets may be single organisations, or they may be a community of interest or domain of specialism such as a sector regulator or industry association. NDT Nodes may be established by individual organisations, regulators and industry associations, or service providers and will be able to handle Digital Twins on behalf of their constituent organisations and provide a secure sharing boundary. The recommended architectural pattern described here is scalable so sharing of datasets is possible both at a sector level between sector members and a regulator or industry association, or at a National level across sectors. It can also be deployed within a single organisation to support internal data sharing, in addition to being deployable quickly for short term use cases such as emergency management response (see Figure 1).

The pattern defines architectural components that will allow datasets to be shared locally (i.e., within an NDT node for a community of interest or even within a single organisation), but also with the functionality to allow for inter-node discovery, authorisation and data sharing to take place (see Figure 1). The interface requirements and standards to ensure interoperability between the architecture components are discussed, along with considerations about how the architecture will operate.

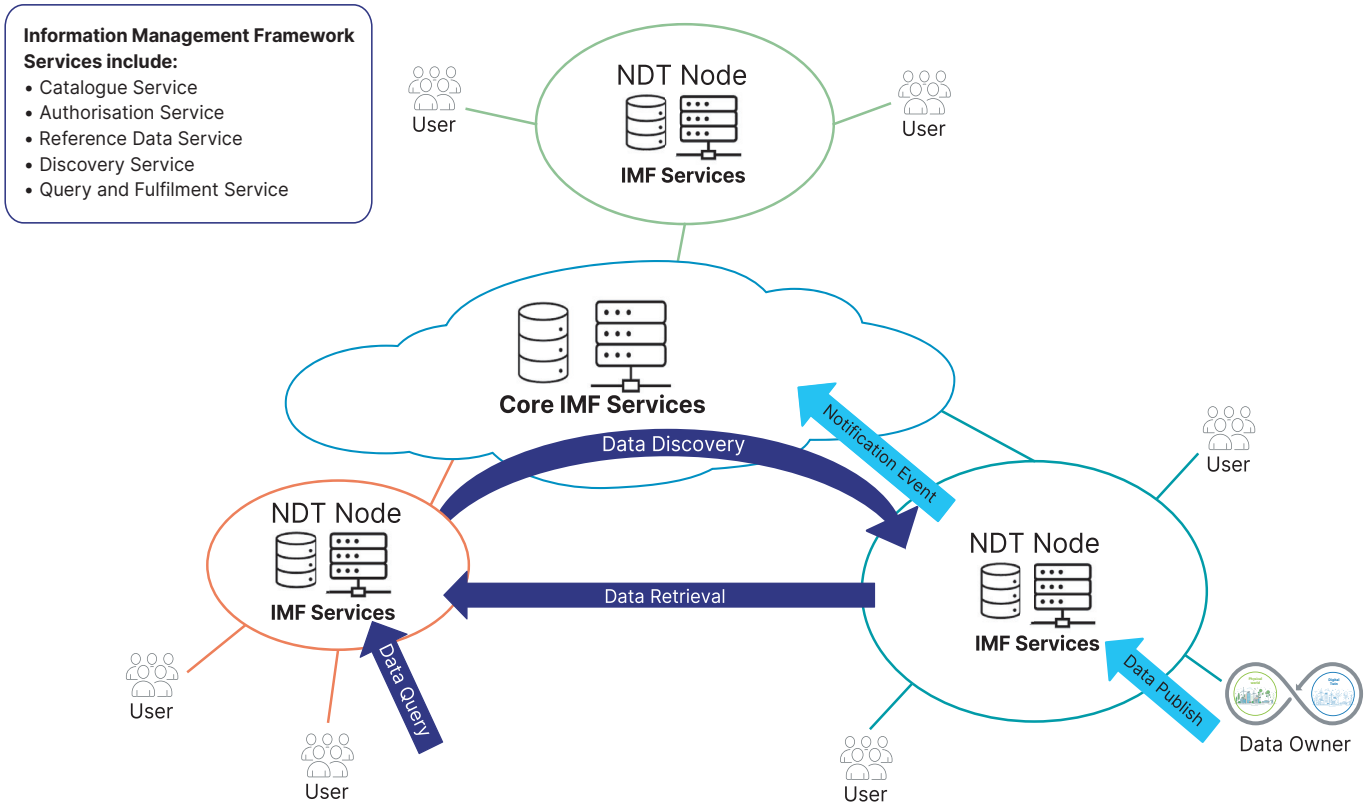


Figure 1: The Integration Architecture Concept

It will provide the required security to ensure that publishers' information is released to and accessed by those parties that have the appropriate authorisations in place.

The Core Services are likely to be quite thin, comprising mainly of: a master NDT Catalogue that holds the location of available NDT Datasets across the ecosystem; the master FDM/RDL that will synchronise with the subset that is relevant for each NDT Node. Propagation of data changes among any Digital Twin replicas will be carried out by streaming the deltas using a publish/subscribe model to parties that have an interest and appropriate contract in place.

The recommended architectural pattern is designed in such a way that it does not prescribe any particular implementation approach, merely that there are NDT Nodes that provide an integration environment for their users and which conform to the external interface requirements and joining rules to participate in the NDT. Once part of the NDT ecosystem, authorised users can share information as their business requires.

NDT Nodes can be established in a flexible way by individual organisations, communities of interest such as regulators or industry associations, or by independent service providers providing a commercial service. It is a requirement of participation in the NDT that all NDT nodes use the same architectural solution, though different NDT nodes will not be required to use the same software solution. Being clear about standard approaches to services and interface specifications mean that it can be deployed flexibly. The only difference is the scaling of the deployment. As long as the principles and 'joining rules' are adhered to, the architecture is agnostic of technology and the stack can be used to suit individual requirements or to leverage existing systems and infrastructure.

Introduction

Purpose of this document

The purpose of this document is to define the patterns, principles and standards needed for the Integration Architecture for the National Digital Twin Programme. The requirement for the architecture is specified in the Pathway report.

The architectural principles laid out in this document define the high-level patterns required to realise the conceptual architecture defined in the Pathway report. This includes the common components and functionality required for interaction between NDT stakeholders. This document also looks at applicable architectural patterns that could be deployed in the NDT.

Although the immediate use case of the architecture is the management of information within the NDT context, it has wider applicability as a general data sharing platform, the core of which is a common data model supported by reference data libraries to enable sharing of consistent data.

Introduction to the National Digital Twin

The National Digital Twin is an ecosystem of digital twins and the protocols by which they can be integrated securely and resiliently. This represents an exciting vision for the built and natural environment.

As set out by the National Infrastructure Commission's Data for the Public Good report in late 2017, a National Digital Twin could provide insights that enable improved operations and maintenance, investment and/or changes to increase infrastructure resilience, reduce disruption and delays, optimise our use of resources and boost quality of life for citizens. This will be achieved by better understanding the performance of national infrastructure and the potential effects of changes to operations and maintenance, and changes to our physical environment before disruptive interventions are made, as will linking between the legacy approaches of different organisations. Further information can be found in the Pathway report.

Requirement overview

The requirement for the Integration Architecture is set out in the Pathway report along with use cases that demonstrate practical examples. An analysis of the document along with stakeholder interviews has resulted in a set of high-level key requirements that are listed in Appendix A. The objective is for the Integration Architecture to be a very general data sharing platform, where using a common Foundation Data Model and Reference Data Library is at the core, enabling consistent data. The Integration Architecture is largely about how to share consistent data with appropriate security. The requirements architecture (taken from The Pathway report) is shown in Figure 2.

The scope for the information to be managed by the Integration Architecture to support the NDT includes all infrastructure that the UK Government owns, operates, or regulates, as well as the services they deliver. In addition, it will also cover all infrastructure assets considered of critical national importance. An asset is defined as a subject of interest to the NDT, it could be anything from a single artefact to collection of infrastructures such as buildings, facilities, telecoms, roads, rail. The nature and scale of the assets will be determined by the participating organisations and the granularity of the asset datasets may differ by use case.

Three specific examples of practical usage as given in the Pathway report are:

1. Complex decision support for retrofit of at-risk buildings.
2. Regional resilience, response, and simulation.
3. The citizen-centric investment into infrastructure.

The time constant of the data/system represented and hence frequency of data update is dependent on the type of the information item and the use case if different protocols and mechanisms are used, in particular where there are break points between different delivery options (streams versus versions of data sets for example). Similarly, the requirements for security and performance will need to be defined to satisfy all uses cases.

For the purposes of this document though, we are looking for general cases, where data from a diverse set of organisations can be shared and analysed for a single purpose or using diverse information from various sources that overlaps to get a bigger picture. It is envisaged that information will be published by organisations:

- To provide information to customers,
- To collaborate with business partners and industry associations, and
- To support legal and regulatory requirements.

In addition to longer term routine operational use cases, an important use case is the ability to respond to an emergency that will

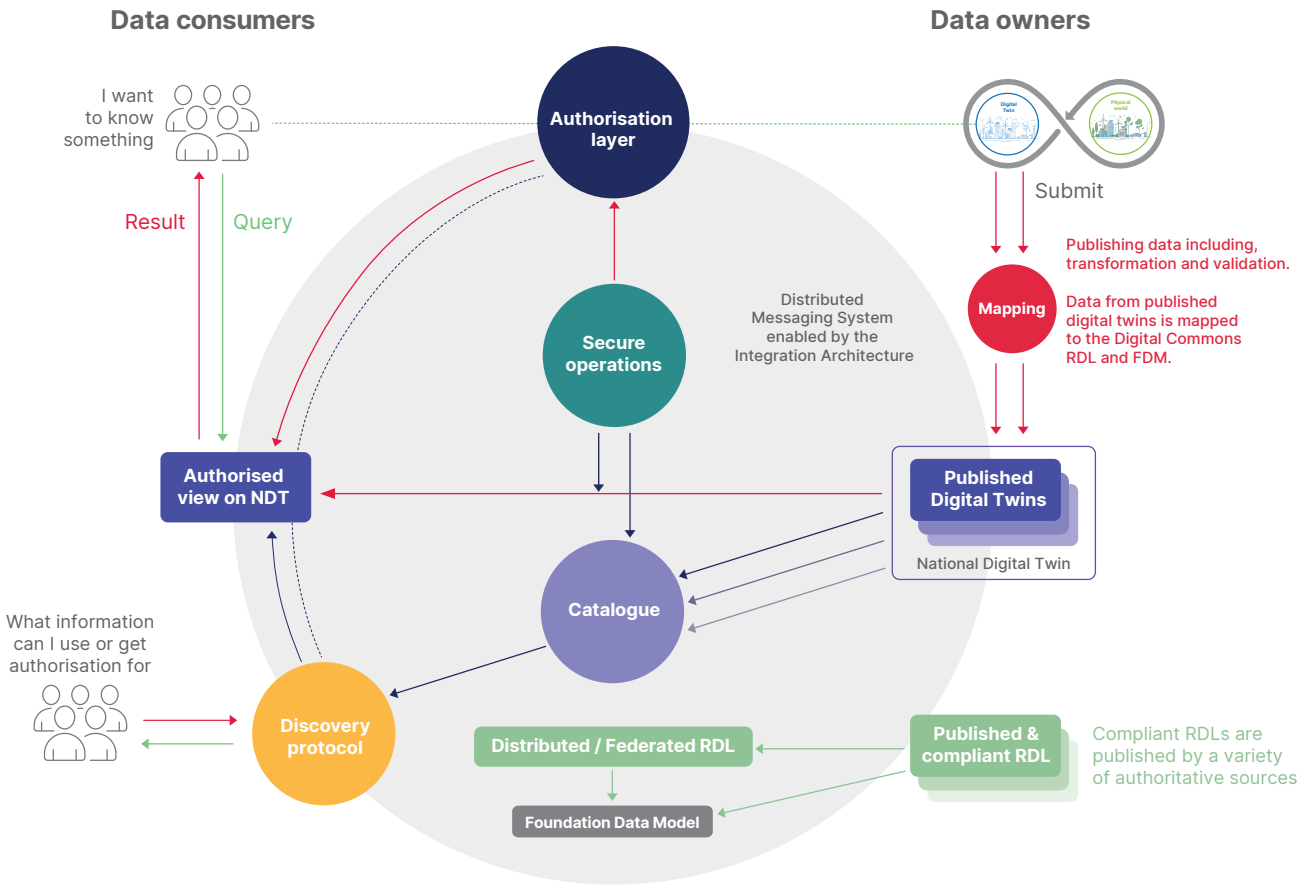


Figure 2: Conceptual View of the NDT – A National Digital Twin enabled by an Information Management Framework

require pulling together data from across different communities in a way that was not foreseen before the incident that caused the requirement, e.g., COVID19. The pattern with emphasis on shared FDM/RDL and interoperability standards is important in ensuring that this is feasible.

In addition to domain operational use cases, it is important that the architecture core functionality supports 'business as usual' NDT maintenance use cases such as:

- Enrolling a new data publisher.
- Find a record outside in another NDT Node.
  - e.g., a regulator wants to locate and have access to data published by another Node.
- Publish a Digital Twin
  - A new asset Digital Twin is released. How does that get published to the local NDT Node, and propagated to other NDT Nodes where copies are held.
- Add a user to the NDT ecosystem – including access rights and propagation of trust.
- Notify of change to Digital Twin.
  - How the pub/sub model is used to register interest, against what topic (taxonomy of concepts).
  - If an Event Sourcing architecture is used (a derivative of pub/sub) then where are the events stored, and who has access?

Trust between parties engaged in sharing is an important factor to ensure that the necessary level of information sharing can take place. Where necessary, trust should be managed by electronic contracts between them. In order to provide the necessary level of assurance, logging and auditing of data access and usage will play an important part. Emerging technologies such as crypto-shredding[1], and potentially D-Apps[2] may be useful to consider for this purpose.

Information attributes that are important for architectural design include accuracy, timeliness, completeness, and provenance that are determined when the data is created; and properties like availability, clarity, and consistency that need to be determined in advance. In addition, the scaling and performance of any architecture will depend on the nature of the data that needs to be integrated and the use case being delivered. These are therefore not specified here.

The need to be able to uniquely identify all items within the Integration Architecture is a necessary function, as will be the ability to identify an existing item on update, so as not to create duplicates. This is particularly important when the update may be a considerable amount of time later e.g., months or years when ownership of assets and/or host systems may have changed.

A core function should be the ability to deduplicate items. In addition, privacy and IPR requirements may mean that de-ID/ re-ID functionality is required in order to allow data to be used for analysis purposes without revealing details that are sensitive. More generally, entity resolution[3] technology may play an important role in statistical de-duplication of data.

Architectural principles

Distilled from the requirements in Appendix A, the following are some key architectural principles that must be adhered to, regardless of the type of architecture that is implemented.

Open data standards

There is a clear theme, running through all the NDT reports and presentations, that data standards are the key enabler to this programme. Work has already begun on defining the upper ontology and data models. Any architecture that is implemented for sharing digital twin data will have to be capable of working with this kind of data. Given that the current modelling work is building on previous 4D ontology work, the building and infrastructure configuration data is likely to be highly connected and graph-like.

However, the configuration data is only one part of a digital twin – there is also data that streams from sensors, often in huge quantities, that is likely to be conformant to a variety of different data standards. It is also likely that this data will often be produced with little in the way of context – i.e., receiving parties need to know where it came from and what it means. It is possible that templates might be appropriate as an approach (e.g.as is done in ISO 15926), where data structure is defined in terms of a mapping to the FDM/RDL, but perhaps the data structure is not transformed.

This variety in the nature, volume and shape of the data means that communities of interest will likely each work with their own subsets of the overall RDL. Care will need to be taken to ensure that the RDL remains a single integrated whole, even though it may be distributed in authorship and publication by authoritative sources. Appropriate governance will be required as well as technological capability.

Data quality

The data models and ontologies being developed for the NDT are likely to expose data quality problems in existing building and construction information. A key requirement for any data that is to be represented in this way is that it must meet a certain threshold of data quality. Major data quality problems need to be tackled at the source; it may be possible to correct minor syntactic errors automatically. In both cases, it is vital that the data quality is measurable and published with the data itself. It is also important that the NDT data sharing implementations provide a standard way to measure and report on data quality, recognising that data quality drifts over time – and usually towards reduced quality. Early detection of problems is preferable to post-incident corrections, so data architects should consider continuous monitoring of data quality and quality improvement by root cause analysis and preventive action by improving processes.

In addition to the question of accuracy in data quality (i.e. – is it correct) there is also the question of currency – i.e., it was correct when it was released but is

now out of date. This relies on the timely publication of the up-to-date datasets. E.g., If the owner/operator does not republish after maintenance or upgrade of the asset, the version available is out of date. Not to mention the difficulties that stem from the differences between 'as-designed' and 'as-built' datasets.

Privacy

Privacy of the published data is a key principle. The Integration Architecture shall ensure that data is shared, and its intended usage guaranteed only according to the conditions under which it was published. Attribute based access control (see section on Attribute-Based Access Control below) is important, as well as limiting the conditions under which the data can be used – i.e., online only or local copy. Such controls must be built into the system and standards identified for how this will be implemented across the NDT ecosystem.

In a distributed system such as the NDT, there has to be a way to share access control groups across organisations. For that reason, we are recommending an attribute-based access control (ABAC) approach. The US Government Enterprise Data Header specification (and the UK equivalent) may be a good way to tag this kind of access control meta-data to NDT data packages[4].

Capacity and throughput

The NDT will contain a diverse range of data and data types, from detailed 3D geometric rendered models of the assets, to streaming operational data from sensors in the assets. These represent two ends of the spectrum for data, large record but largely static data through to small packets of real time or near real time data. This diversity means that the demands on the system will vary with use case. Any implementation of a Digital Twin will have to be able to be designed with a view to performance and scaling that suit a particular data type. There will inevitably have to be trade-offs made regarding the performance and scaling, and the users within the community of interest will have to tailor their use cases accordingly. It is also the reason that a given technical solution to this architectural pattern may not be suitable for all use-cases – hence the emphasis on standards and architectural principles rather than mandating specific tools.

Compliance and deletion

Information governance rules will need to be compiled (on a domain or use case basis – See Next steps) to ensure that data is coherent, current, and only accessed and used by authorised parties under specified conditions (digital contract). Data owners need to be assured that their data has been accessed, used, and deleted by parties it has been shared with. Compliance with any terms will be carried out by automated rules-based components.

Security

The architecture shall ensure that all data and functions are secure from bad actors. Authentication shall be distributed, but trust shall be propagated throughout the system using trusted digital ID providers. In addition, data at rest and data in transit security standards shall be applied according to best practice and shall be updated in line with advancements in the area. As a general architectural principle, a zero-trust approach should be adopted, in line with NCSC guidance[5].

In some specialist cases around critical national infrastructure, the data being shared may be protectively marked to such a level that they can only be exchanged between HMG approved systems. This will of course require adequately protected network endpoints and sovereign cryptographic systems. However, it is likely that protectively marked and open data will need to be integrated and handled together, so the security implementation of data at rest and data in transit controls needs to be able to take this into account within a single overall architecture that understands different security levels.

Data integration and coherence

The published data shall be defined and structured using the FDM and RDL. It shall be uniquely identified to ensure that integration can be achieved across all Nodes of National Digital Twin. The unique identification shall ensure that updated elements are correctly identified and prevent the creation of duplicated elements (See Next steps). In order to minimise the Core functions in the NDT Integration Architecture, a policy of using a non-managed mechanism such as Version4 Universally Unique Identifier (UUID)[6] would seem to offer the most benefits as there is no assignment authority required for a namespace. In addition, most platforms offer a UUID generation service.

Attribute-based access control

Attribute based access control (ABAC) allows more granular access to data based on matching attributes on the data with allowable values for the users[7]. The range of information to be accessed, and the potential uses that it will be put to by a wide variety of users suggest that simple role-based access will not be sufficient. We can therefore prevent access to sensitive information based on a security classification; or limit access data about a particular context of relevance to the user (e.g., oil and gas, or safety certificates, or building regulations). If the attributes carried by the user match those on the asset dataset and according to the contractual terms in place, then access is allowed.

What will be required is wide agreement on the schema or taxonomy of attributes, and a service that manages the access controls and provides authorisation for access to the data (See Next steps). Work is in progress toward identifying the possible mechanisms for implementing such controls by the National Digital Twin programme.

Encryption

Encryption will be a key aspect of the security features in the NDT. Data concerning aspects of the UK's critical infrastructure will present a tempting target for any bad actors, and the consequences could be significant. Best practice standards shall be used for data at rest and data in transit. For data at rest total disk encryption is required, and is commercially mature technology, whilst for data in transit public-private key cryptography is recommended. For instance, the Transport Layer Security (TLS) V1.3 is current best practice[8].In both these cases it is important to have a key management service in place. There are many commercial solutions available as well as built in services to most cloud platforms (e.g., AWS Key Management Service[9] and Microsoft's Azure Key Vault[10]).

Secure deletion techniques (e.g., Crypto-shredding[11]) may also be a useful technology for the NDT and should be used where appropriate.

Open source

In order to make the barrier to entry to The National Digital Twin programme it seems appropriate that implementations leverage as much existing open-source technology as possible. It is also important that any HMG funded work is at least released under a license that is permissive to HMG. This could mean open source in many cases (e.g., the Open Government License – OGL is already UK policy). In some cases, something like MOD's DEFCON705 may be appropriate – e.g., where HMG is keen for British companies to exploit their technology profitably outside the UK.

Further phases of this work will need to spot gaps in the required standard and form recommendations for work to create open-source solutions (see Next steps).

Standards-first

The data sharing environment for the National Digital Twin will require on-boarding of disparate stakeholders, with different IT stacks in their own enterprises, and varying levels of data maturity. Employing recognised standards wherever possible will be essential to uptake. The joining rules must be balanced to ensure that there is a low enough barrier to entry so as not to inhibit take-up, but high enough that data quality, security and compliance are not detrimentally affected.

A candidate standard amongst many is ISO19650 [ISO, 2018][12], however this will need upgrading to make it appropriate for use in the NDT, and in addition there will be a need to develop standards where they are missing. This will be a key part of the NDT programme.

Cloud-ready

Although there will be exchange at the enterprise boundary, it is likely that significant use will be made of the major cloud providers, and this is only going to increase over time. To maximise integration and interoperability opportunities, any architectural choices should be dictated by the question “will this work in a cloud environment?”

Architecture components

Each of the Architectural Options and Patterns that follow are composed of functional components that are required to deliver the Integration Architecture vision. This section describes the components; however, the reader is directed at the sections that follow to understand how they are deployed and how they interact in each option. The scale of each architecture shall be dependent on the local usage and conditions. It is envisaged that the components shall be composed of both proprietary products where necessary for legacy systems, but also that a community of open-source modules that adhere to the architecture principles will be created for wider community use. The important aspect is that they comply with the required standards to ensure interoperability. For each use case there may well be performance and other constraints that will determine the actual choice of tools to fulfil the functionality described here.

Authorisation

The authorisation service uses trusted digital ID providers (for example those organisations that provide identity for GOV. UK Verify) to provide the credentials that can be used throughout the ecosystem, as long as the node in question has agreed the source as trusted. These digital ID providers (Post Office, Experion etc.) use a variety of information sources to verify the identity of an individual and create a digital identity that can then be used (with authorisation) to access digital services.

For example, Barclays trusts Apple to issue a verified digital identity to an individual, this is then stored on their mobile device. The digital identity is then authorised for each use using a biometric mechanism (e.g., fingerprint or face scan). Barclays trusts this digital identity issued by Apple to allow access to personal and financial information stored within the bank's own systems. In plain terms it's possible to view my bank account statement in an app on my phone because Barclays trusts that it's me who is using the phone.

This propagation of trust negates the requirement for a central identity storage capability. Authentication can be provided by any of the very many authentication services available (e.g., Thales SafeNet, Microsoft, Amazon, Google, OneLogin). To ensure inter-community interoperability there shall have to be agreement on the attributes that are required as part of the credential to allow the attribute-based access to data. Like the model for the RDL this should be a harmonised set and the subset that is required for each digital twin propagated out.

Service providers usually support recognised standards in this area such as OAuth 2.0[13] for authentication and OpenID[14] for Identity services and Security Assertion Mark-up Language (SAML)[15] for both.

Reference data libraries

The reference data library holds and allows for querying and delivery of the reference data from repositories using the Reference Data Service. The local RDL would provide a relevant subset of reference data, synchronised from the Core RDL for the relevant community or organisation. Community members, e.g., Equipment manufacturers, might themselves be authoritative sources for data about their products. There are authoritative sources such as BIPM who are the OEM for reference data they are the source for. This will likely be provided to the NDT in the same way as a Digital Twin (i.e., mapped into the FDM/RDL from the original).

The synchronisation of the local Reference Data Libraries with the Core RDL shall occur using a publish/subscribe approach as a streaming based service or by delivery of the Core RDL in a container that can be integrated with the local RDL.

NDT catalogue

It provides a catalogue for digital twin datasets. This contains the location of available Local Directory Managers at a minimum. It may also contain the owning repository and routing information required to access the NDT datasets. This will consist of location and identifier of the repository and an unambiguous identifier for the datasets (see Section Data Integration and coherence above). It should also contain the access and usage information with which the dataset was published which can be queried during dataset discovery. There will be synchronisation of the NDT Catalogue to the Local Directory Managers which shall act as local caches in order to minimise traffic to the core and improve efficiency. Much in the same way as DNS Servers are cached locally for resolving internet addresses more quickly.

Local directory manager

The Local Directory Manager shall manage the NDT catalogue that contains references to all the datasets that has been published to the Integration Architecture at local level. A subset of these shall be tagged for wider access and therefore the directory information shall be propagated to the core catalogue. The local directories shall also have to be able to synchronise with other local directories on a per contract basis in order that datasets can be discovered outside of their own domain. The synchronisation shall be carried out using event sourcing such that all changes are stored as a sequence of events, with the deltas available for permitted subscribers on a publish/subscribe basis.

Information management service

The information management service is the key component for the Integration Architecture pattern. It acts as the interface for all discovery, query, publication, and query fulfilment. It also acts as the primary component for management of the community of interest and management and

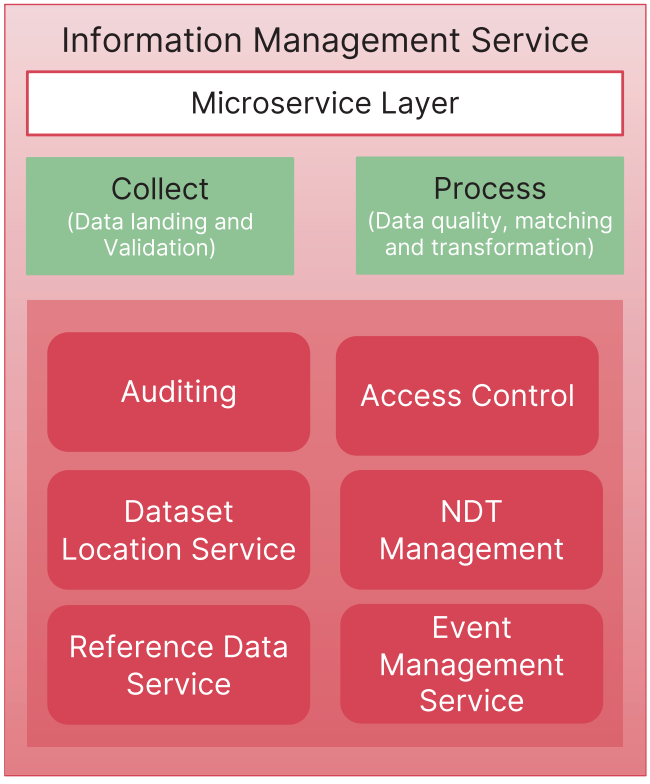


Figure 3: Information Management Service components

integration with other communities of interest. It shall expose microservices that provide the required services. A detailed microservice catalogue will need to be formed in future stages. The information management service will provide the linkage and processing between the requests from the Microservice Layer and the various repositories containing the NDT Catalogue, the RDL, user identities (if present). It will provide the routing information for dataset discovery, query and access and the events required for the notification of the pub/sub model.

Collect and process

The Collect components perform the functions to receive any datasets and validate them against the appropriate schema.

The Process component will perform data quality checks, ensuring internal integrity and integrating against any existing known data (e.g., checking for duplicates that have already been released – we don't want 2 Buckingham Palaces!). It will then perform any transformation necessary to either correct any issues found to put it into a form for onward transmission.

It should be noted that these functions may be a very thin layer or even non-existent, as it is likely that joining rules stipulate the quality thresholds, formats and schemas that have to be adhered to, however there will still need to be some conformance and compliance checking. It is a design and policy decision as to how datasets will be released into the NDT and therefore the amount of checking required. Indeed, if the data is released and queried at a dataset level this will be less of an issue. If data is released at a record level, then arguably more Collect and Process functionality may be required.

Auditing

The discovery and access of the datasets is carried out in a controlled manner according to contracts agreed between the parties. As part of the overall security mechanisms and for forensic analysis after security events, all interactions shall be logged, and auditing capabilities provided to allow investigators to have access to data that can identify root causes and perform remedial action.

The auditing shall also have the capabilities to monitor system performance and adjust scaling if necessary. It can also be used (with appropriate de-id functionality) for analysis and to gain insight into the use of the NDT.

Access control

Although there are the opportunities to use external identity and authentication services, the management of the validated credentials particularly for the access of datasets across communities of interest will be important. In particular the management of the attributes to be used in the attribute-based access, validating the identity and assigning attribute values and access rights to the digital identity. These can then be used to discover, access, and publish datasets appropriately.

Dataset location service

Using the NDT Catalogue the Data Location Service will provide the routing to authorised users that are discovering, querying, and accessing datasets. To reduce hits on the NDT Catalogue and improve responsiveness, it may need to maintain a cache of available NDT repositories which will be used within its own Node. The Core NDT Catalogue will allow Dataset Location Services to identify repositories and datasets outside the local Node.

NDT management

There are a number of management functions related to the operation of the NDT. Functions such as managing the access model, adding, and removing organisations to the NDT. Setting up routing to repositories, catalogues etc. The exact range of functions will depend on the final decision as to how much functionality is distributed, but a core set that rely on interaction with other functional components shall be identified as part of the Microservice definition task.

Event management service

The notification of changes (e.g., publication of datasets) will be carried out by notification using an event management service. The event management service shall operate on a publish/subscribe basis, with messages delivered to subscribers of matching topics. The range of messages and the taxonomy of topics shall need to be determined but will include items such as:

- Domain sector (electricity supply, rail, health, architecture)
- Regulation and planning
- Safety certification
- Faults
- Hazards
- Emergency response

Microservice layer

The Microservice layer will provide the main interface to all IMF services. The core and each 'node' either a community of interest or individual organisation that wants to participate in the NDT shall interact through the microservices, with which they shall have to be compliant. The individual services will be implemented using a number of standards and technologies, whatever is appropriate for the service to be provided/ consumed e.g., streaming services, synchronous or asynchronous data transfer, authentication, or data querying. The definition of the services later will be a major work item required in the phases to come (See Next steps).

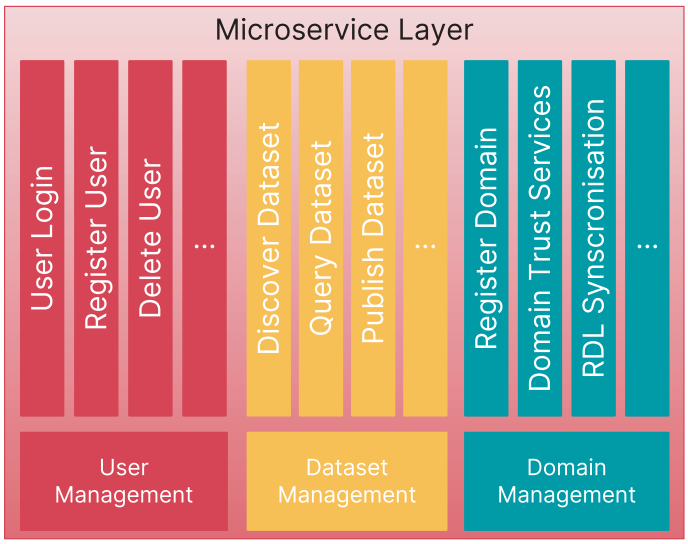


Figure 4: Microservice Layer with example services

Integration architecture options

This architecture options section describes three high level architectural patterns that are viable candidates for the NDT Integration Architecture. These patterns are a centralised, federated, and distributed architecture. They represent the endpoints and a mid-point in the spectrum of systems integration architectures and are all feasible to different degrees to realise the NDT Integration Architecture.

The main functional components required by the architecture and shown in the diagrams are explained in the previous section. They are used in the patterns that follow; however, the configuration, scale and specific interfaces will vary across the use case.

Pattern 1: Centralised

In a centralised architecture the functions and data are centralised in a single virtual node as shown in Figure 5. The providers of the data and the consumers all sit outside the centralised node and interact with it. This means that there is a single location for all services for all use cases, meaning the synchronisation issues are reduced. Authentication, data ingest, transformation and storage, and other management functions all occur with the single node. There are no issues with maintaining pointers to data location as its all centralised. However, this approach has a number of drawbacks, mainly in that the provider organisations either have to keep the information within their technical and governance boundary or release it to the centralised core node. There are limited opportunities for sharing between limited number of participants in a domain boundary or for communities of interest to share RDLs relevant to their domain. All services are reliant on a single core node that will have to scale to manage the super set of all participating organisations. Although virtualised infrastructure with scalable load balancing can be suitable for this model, it may produce performance issues depending on the number of participants and the nature of the datasets and the operational usage.

Key benefits:

- Standardised approach to managing NDT data, based on a set of agreed guidelines and policies across all organisations. Therefore, enabling a consistent level of assurance.
- Authentication and other management functions are consolidated in the one node there is no need to synchronise identities or other data across different organisations.
- There is an identifiable, reconciled, verified source for the NDT datasets that is maintained and that can be referenced and trusted.

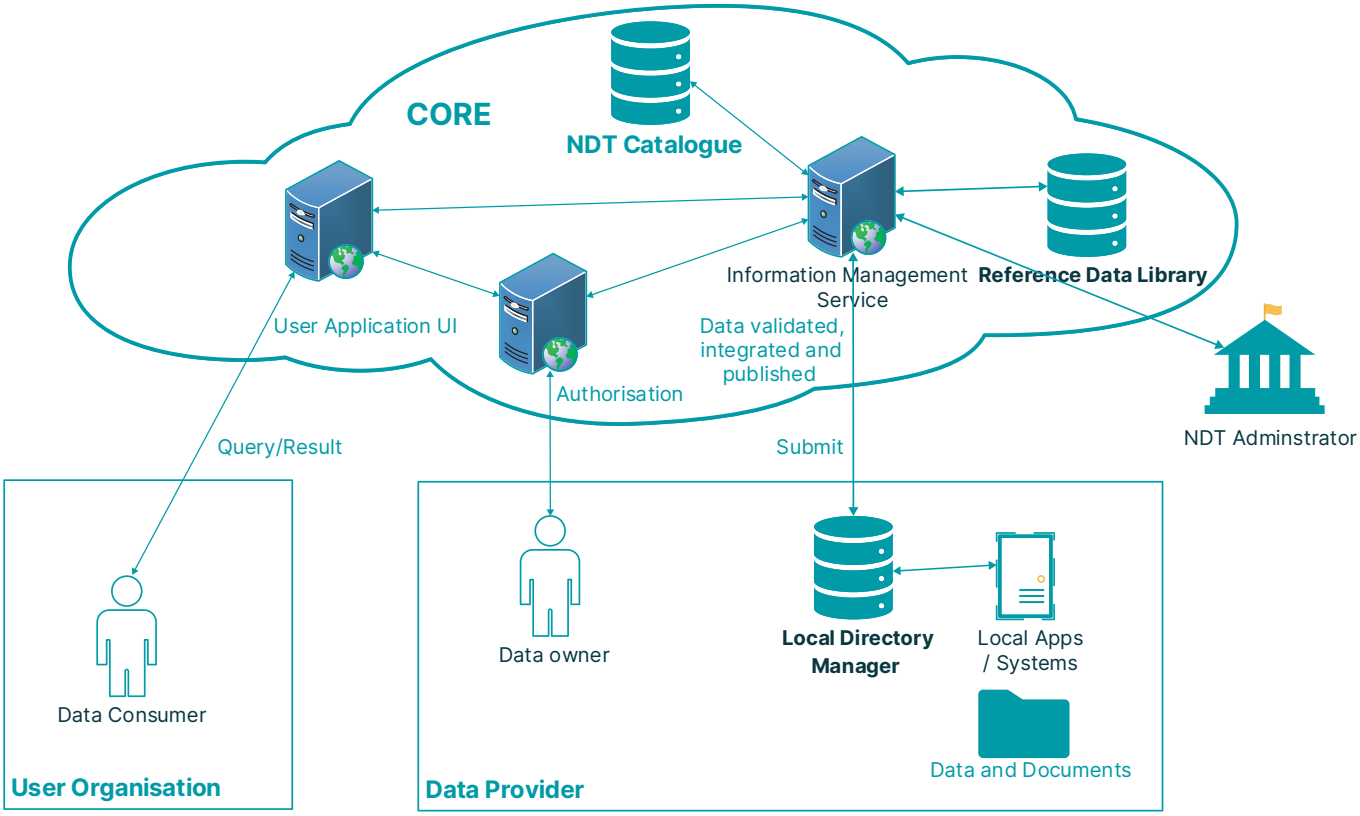


Figure 5: Centralised Architecture Pattern

Key concerns:

- Typically, a low level of owners control over their data. The model has an inherent dependency on the centralised authority ensuring appropriate maintenance, use, update, and revocation of data.
- There is a high level of dependency and risk on the centralised Infrastructure as a single point of failure.
- The model creates centralised data repositories of catalogues, authentication credentials etc., which may present an attractive target for data breaches. However, it should be noted that the centralized model does greatly decrease the attack surface. If there are dozens of nodes, each controlled by organisations with varying security skills, then one weak one could open them all up to abuse.
- Careful consideration will have to be given to ensure that a centralised system architecture could scale to meet the needs, of sometimes diverse, sets of domains.
- The centralised access control model may not suit the needs of the individual domains which might require access control more locally using RDLs based on their own requirements.

For the reasons given above the Pathway report has already discounted this option so this is unlikely to be the way forward.

Pattern 2: Federated

In the Federated Model data and functions are generally provided by different stand-alone systems with a common trust and interoperability framework (consisting of data standards, communication protocols, access management standards etc.) rather

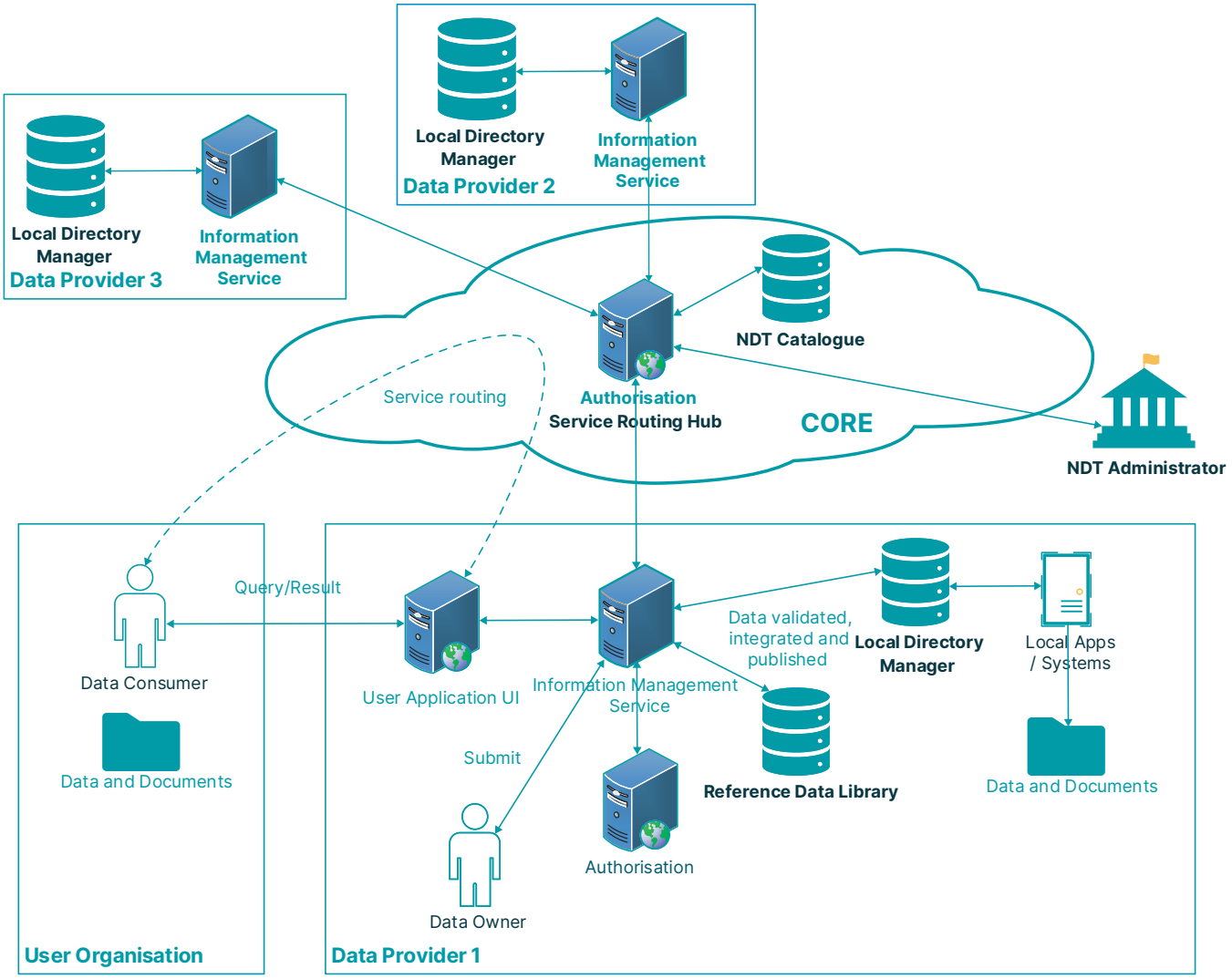


Figure 6: Federated NDT Architecture Pattern

than a centralised, single point of integration. This allows interoperability with services that comply with the framework, that may be within the NDT or other third-party services (e.g., identity management, data analysis/ science). However, some functions may be centralised depending on the requirement (e.g., identity management).

An example shown in Figure 6 each provider organisation maintains a repository and directory of NDT data, which will be published to the single Federated Information System. Most functionality and data is devolved to the local organisations with requests and responses from users managed by a Service Routing Hub using the NDT Directory architecture component providing the definition of available service endpoints (e.g. dataset repositories) and a routing service as shown by the dotted line in the figure. The Service endpoint discovery is based on Provider Organisation ID, and a variety of other attributes that classify the data that they have published. The Service Routing Hub maintains a record of all data provider organisations URLs to route requests accordingly. Requester's Organisation ID and Roles are used for authorisation, so long as the user has a contact with the provider organisation.

Key benefits:

- In some cases, a federated model can achieve high levels of adoption due to the decoupled nature of such a system. In effect, the controls applied are through technical and data standards as opposed to a mandated IT infrastructure.
- Can plug into the more general digital ecosystem in the UK driven by the provision of online services both by the government and major institutions such as utilities, banks etc. In this way individuals can bring verified credentials from third-party organisations thereby potentially reducing operational cost of running authorisation services and provide access to a wider range of data.
- The high level of interoperability supports the management of data across different technical systems operated by provider organisations.
- Since all functions are loosely coupled, this pattern allows for a competitive ecosystem of verification, storage, and other service providers to be selected by provider organisations, subject to qualification to minimum criteria and adherence to the joining rules.

Key concerns:

- The contract basis for access to datasets mean that there is a high level of complexity associated with creating multiple one-to-one trust relationships (N x M) rather than via a central authority (core or domain), which can inhibit wide scale implementation.
- There is an onus on local organisations / NDT administrator to agree multiple legal and technical agreements to enable a federated relationship. This has an associated level of complexity and therefore also cost. This would need to be mitigated by specifically looking at legal issues and standard agreements that could be used widely (See Next steps).
- Administration of the data control and routing functions sits with multiple parties (i.e., federated authorities) and could present an increased risk of compromise if they all have varying security skills, then one weak one could open them all up to abuse, the security and trust model needs to ensure that this is not possible.
- In a truly federated system, where all functionality is devolved out from the centre, the lack of a centralised system could mean that in cases of conflict between data attributes obtained from different sources (e.g., from the builder and the owner/operators). There could be issues with local organisations interpreting these results differently. Ultimately, this could lead to a lack of trust in the model. This can be mitigated by ensuring that there is a strictly enforced data ownership model, with clearly identified authoritative sources.

Pattern 3: Distributed (or peer to peer)

Different stand-alone nodes with all or most functions localised. Minimal coordinating functions carried out centrally. Even then, the centrally held data may be replicated locally for performance reasons.

A distributed model enables provider organisations to control and manage access to their data. This model comprises an Information Management Service (supplied by a provider – or proxy organisation) and a record locator store, which is also owned by the provider organisation/proxy. The record locator store holds the pointers to the information that is within the ownership of the data owner or their proxy (e.g., a proxy can be delegated to manage information for a specific domain of interest or ecosystem or organisations). The individual organisations can choose who to share their data with, either directly or via any number of proxies.

A common identity and authentication model will ensure consistent application of attributes to user credentials. Access to data will be via attribute-based access (the attribute model to be agreed). Trusted third party providers can also provide authentication

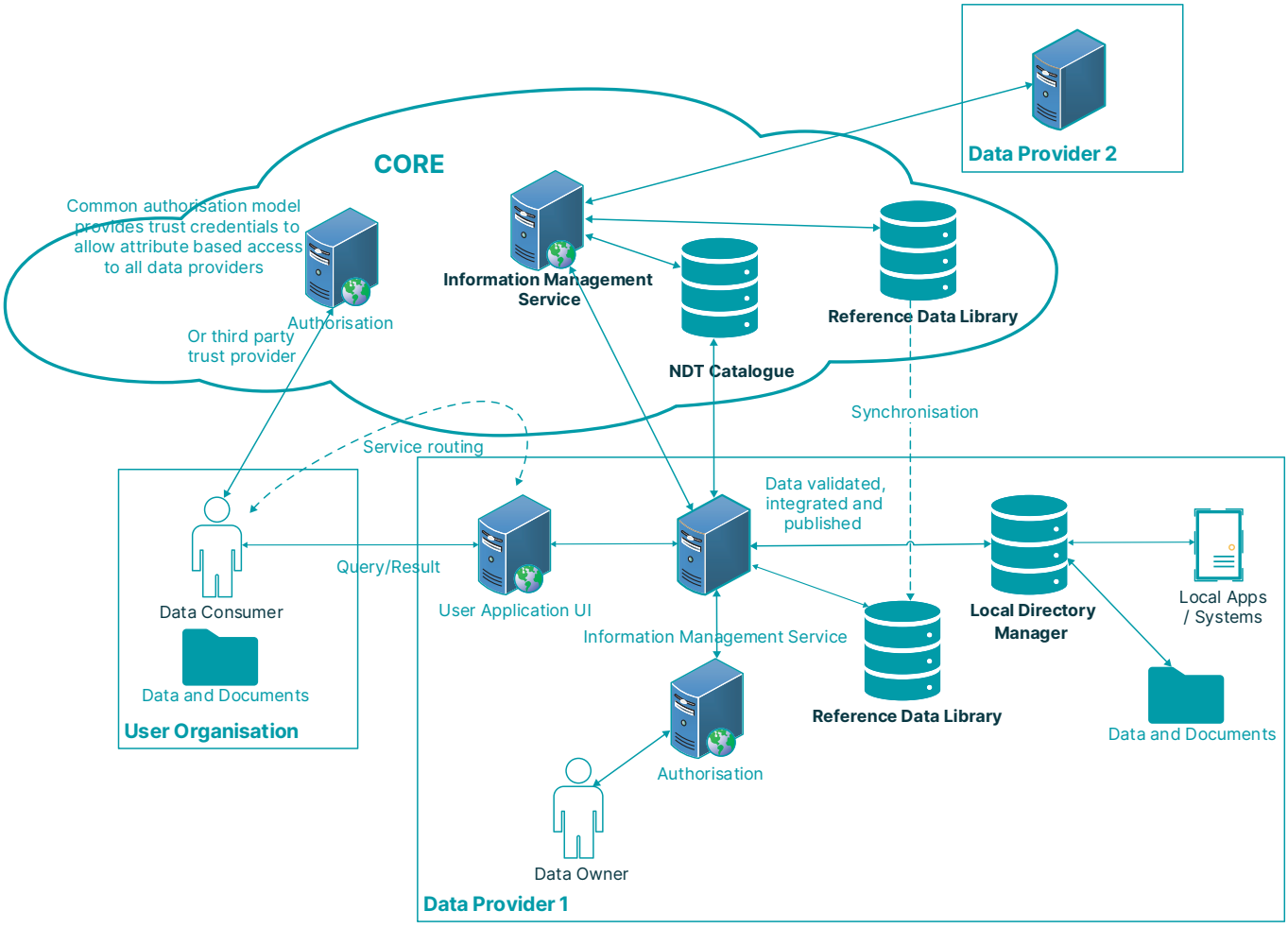


Figure 7: Distributed NDT Architecture Pattern

services, thereby lessening the overhead and cost of providing authentication services. Other centralised services would include the reference data library of core and common concepts and ontology server. A centralised dataset catalogue or at least catalogue of catalogues that knows the location of each of the domain catalogues (akin to a domain name server in the internet analogy) is likely to be needed for discovery purposes. Only those concepts from the FDM/RDL that need to be centralised are done so, the vast majority of the reference data is distributed and based on RDLs provided by authoritative sources.

Key benefits:

- An improved user experience: the distributed nature of the model enables a level of personalisation for each organisation so that the user sees information and concepts that they understand and are relevant to them.
- There is a high level of interoperability.
- Reference data library components can be created and managed by authoritative sources, as long as they conform with the FDM/RDL.

Key concerns:

- Distributed models are low maturity, therefore, there is an element of risk associated with implementing a distributed model at scale.
- There is a dependency on agreeing an attribute-based authentication model.

- A clear governance system for a distributed model will need to be designed (See Next steps).
- It may involve some duplication of functionality at the distributed/local levels, which may have the side benefit of providing some redundancy.

Recommended integration architecture pattern

Given the analysis of the patterns above, the recommended architecture takes characteristics of the Distributed or Federated patterns to produce a flexible solution to meet the requirements. It consists of a re-deployable pattern that can be used at different scales according to the use case. The following represent examples:

- Within a single organisation to do internal integration and interact with the NDT.
- Within an integration environment serving a number of organisations using the same subset of the overall FDM/RDL (e.g., a community of interest).
- By a service provider who wishes to provide Digital Twin publishing/access services to their customers (analogous to website providers).
- Between integration environments with different FDM/RDL (e.g., 3D/4D)

Figure 8 shows the high-level concept of the recommendation. The architecture is deployed to integrate organisations that are part of a community of interest (in fact it can be deployed by single organisations as well). All the communities interact via a microservice layer with the core which is a slimmed down version of the pattern and offers essential core services. Datasets are published by the data owner (1), these are then made available to the organisations within the community of interest, in addition an event is issued to register publication with the core (2). When queries are submitted (A), the dataset can then be discovered by organisations in other communities of interest (B) and retrieved where appropriate (C). The release, discovery and retrieval are carried out according to the authorisation service so that access is controlled as specified by the data owner.

The detailed logical view showing the major interactions between the Architectural Components is shown in Figure 9. This illustrates that the deployable pattern can be used at the Core and Node level – indeed it can be used at the organisation level as well to interact with a Node or with the Core directly if it is not a member of a Community – however this isn't shown in detail. Dotted lines summarise endpoint to end-point interactions that actually take place via the services provided by the Information Management Service.

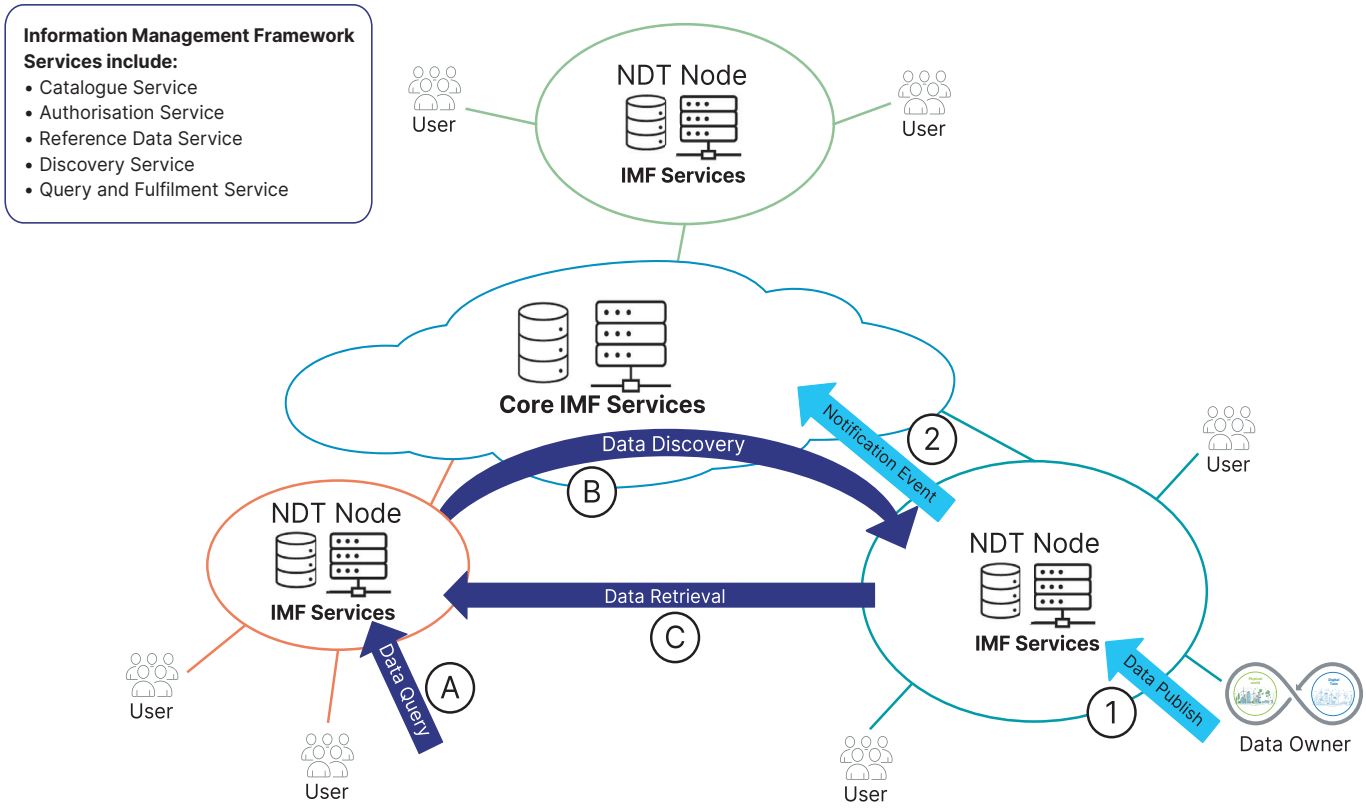


Figure 8: NDT Architecture Concept

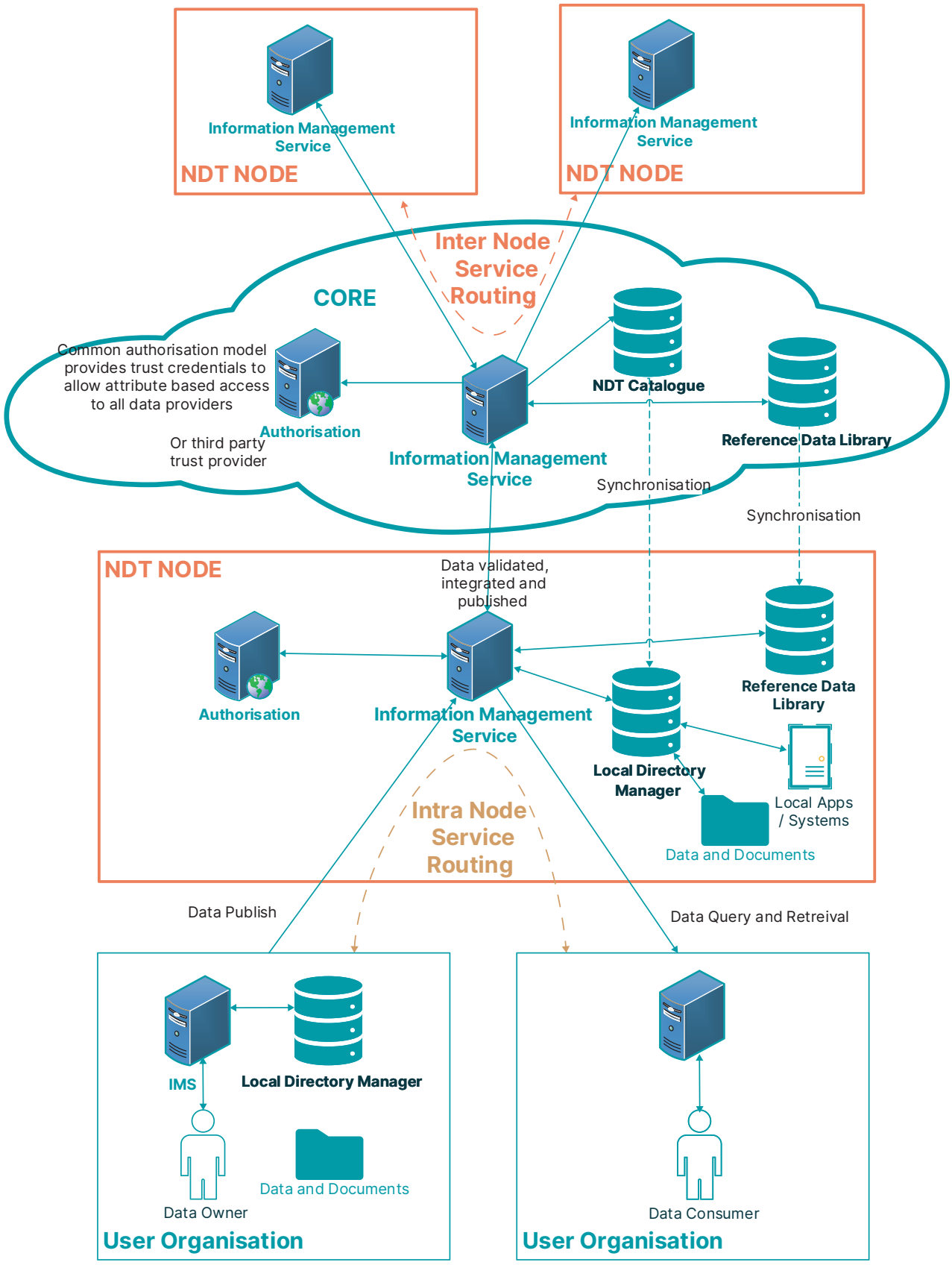


Figure 9: Deployable pattern used at Core and Community of Interest levels

Interfaces

Each of the interactions between the components in the Integration Architecture in Figure 9, represents an exchange of information. This section describes the interactions that are to be used for each of those interactions. The principles and the standards will form the basis for joining rules for organisations to integrate and participate in the NDT. The numbered interactions, in Figure 10, are described in more detail in Table 1 below and the schema requirements to support these interactions are shown in the Data Integration Pattern in Appendix B.

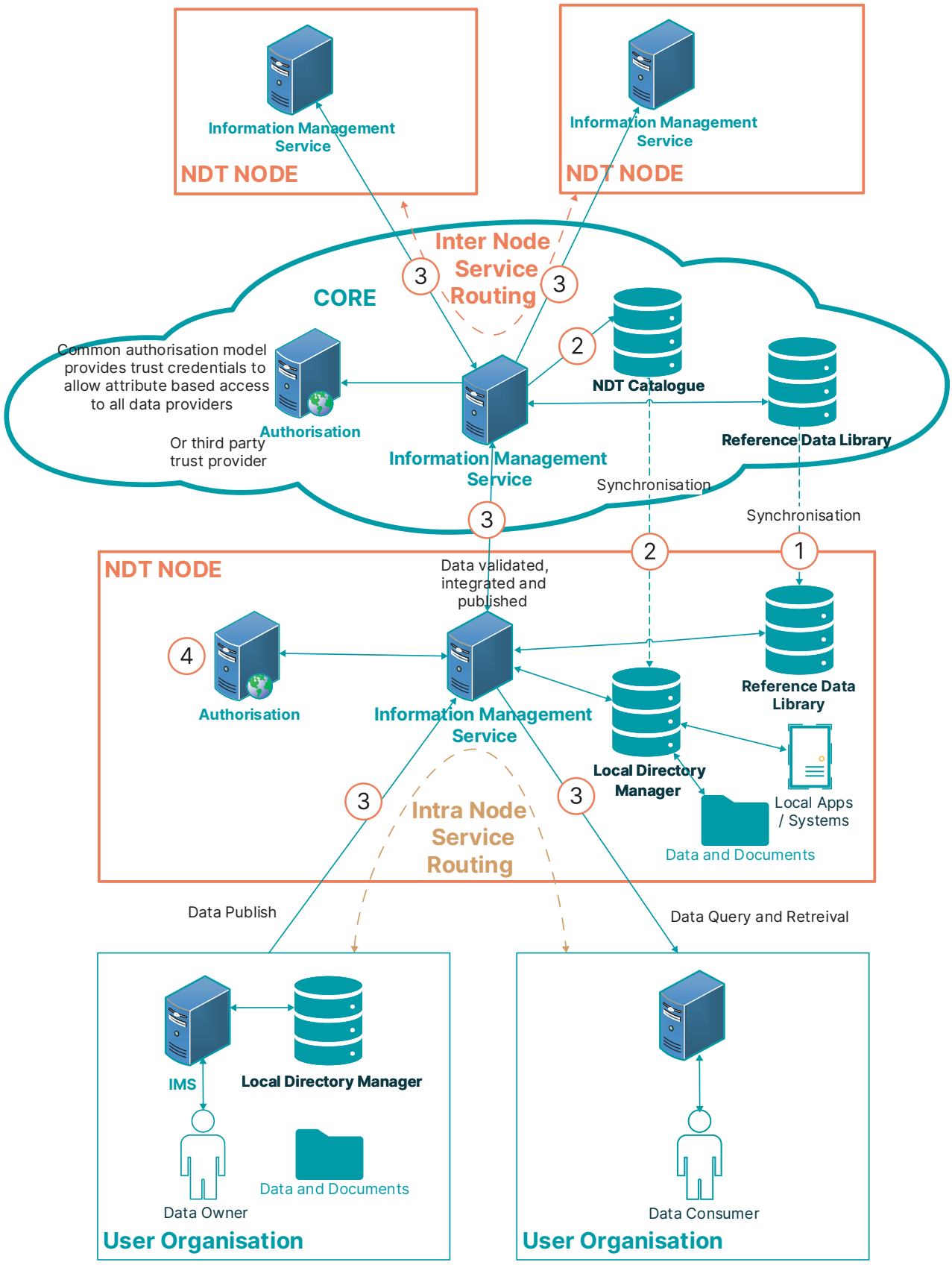


Figure 10: Recommended architecture pattern with interfaces identified

Interface	Source	Target	Description
1	Master RDL repository	Local RDL Repository	Synchronise RDL repositories and propagate throughout the NDT ecosystem. Reference data is updated occasionally, and it is usually new records only (at least it should be). Not all reference data should be synchronised locally, just that which is relevant to the local NDT Node. A mechanism for selecting and pruning reference data will be needed, probably together with publish/subscribe model. Where there are changes in reference data, a managed process will be required to perform the updates, since it could require changes to large numbers of records in source systems to accommodate the change.
2	NDT Directory (Catalogue)	Information Management Services	Provides the information management service with the locations of the NDT data repositories and datasets.
3	Core/NDT Node Information Management Service	Local Information Management Service	To ensure that the Core is aware of NDT Nodes, their capabilities and other attributes. In the recommended model, the core IMS will be thin, mainly being an authoritative source of Reference Data.
4	Data Consumer (User)	Authentication Service	To authenticate a user against trusted credentials and assign the credentials with approved attributes to allow role-based access to the appropriate data.

Table 1. Integration Architecture interface specifications.

Next steps

The list below states some of the Tasks that need to take place to develop further the Integration Architecture components.

Task 1. In conjunction with the wider IMF work and with industry partners, use cases and scenarios need to be developed that allow the requirements and design of the Integration Architecture to be further developed. These will allow detailed examination of how the components in the Integration Architecture interacted and the information flows required.

Task 2. Design the mechanism and process required to propagate out RDL changes locally. Not all reference data should be synchronised, just that which is relevant to the local community of interest. A mechanism for selecting and pruning reference data will be needed, probably together with publish/subscribe model.

Task 3. Create a service catalogue for microservices to be delivered by the Information Management Service.

Task 4. Information governance rules will need to be compiled to define how different types of information is handled and processed. It should include any rules for Personal Identity Information (PII) as well as guidance to comply with security, IPR and legal requirements.

Task 5. Determine how unique identification of items shall be carried out across the entire architecture.

Task 6. Determine the schema or taxonomy of attributes, to provide attribute-based access.

Task 7. Identify gaps in the required standards and form recommendations for work to create any missing standards or open-source solutions.

Task 8. Determine the events required to be handled by the Event Management Service, what their triggers are and their expected payload. Also determine the taxonomy of topics that can be subscribed to by the recipients of the events.

Task 9. An early proof of concept to validate the Integration Architecture components and their interactions. This will provide the basis to start producing open-source components to fulfil the Integration Architecture functions.

Task 10. In order to minimise the risk and administrative overhead of sharing agreements work should be carried out to look at legal issues and standard agreements that could be used widely.

Appendix A
Key architectural requirements

This annex covers the key requirements emerging from the NDT documents and from interviews with key stakeholders. They incorporate concepts from, and are consistent with, the Pathway report including the Gemini Principles (CDBB, 2018). As such these are not a detailed set of technical requirements, but those at a high level that are most important to form logical architectural patterns. Any proposed architecture must be capable of meeting these requirements.

The principles and requirements cover the following categories:

- Architectural principles
- Security/access control
- Interfaces
- Information governance
- Data
- Standards
- Legislation

The table below summarises the requirements that are expanded in the sections that follow.

Requirement	Categorisation	Source
The Integration Architecture shall consist of multiple interoperable and consistent instances of the key functional components.	Architectural Principles	Pathway report 3.7 p30 Confirmed by stakeholder interview
Access to data will be controlled by an authorisation component or layer.	Security	Pathway report 3.7 p30 Confirmed by stakeholder interview
Data owners shall be able to make data visible and available to authorised users	Security	Pathway report 3.7 p30
Actors shall be able to link new and existing twins through the architecture	Data	Pathway report 3.7 p30
Actors shall be able to query twins and assets throughout the wider ecosystem.	Data	Pathway report 3.7 p30

Requirement	Categorisation	Source
The Integration Architecture shall be able to validate published data	Data	Pathway report 3.7 p31
The Integration Architecture shall be able to transform published data	Data	Pathway report 3.7 p31
The asset owner shall be able to publish using a variety of different integration engines	Architectural Principles	Pathway report 3.7 p32
Data owners shall be able to specify the purposes to which their twins or the data produced from them, may be put.	Data/User	Pathway report 3.7 p32
The Integration Architecture shall use the FDM and RDL for the validating the definition of the data.	Data	Pathway report 3.7 p32
The Integration Architecture shall use the definitions specified in the FDM and RDL for allowing access and usage of the data.	Data	Pathway report 3.7 p32
The Integration Architecture shall have protection in place to protect intellectual property and commercially sensitive data.	Security	Pathway report 3.7 p32
The Integration Architecture shall protect the data from malicious or hostile interference.	Security	Pathway report 3.7 p32
The Integration Architecture shall provide a catalogue of digital twins considered within the National digital twin.	Data	Pathway report 3.7 p32
The Integration Architecture shall allow providers to register and specify the information and services that they intend to offer.	Security	Pathway report 3.7 p32
The Integration Architecture shall allow the user to query the catalogue and interact with the data as if it were a single database.	Data	Pathway report 3.7 p32
A messaging system shall allow communication between the digital twins and reference data libraries,	Architectural Principles	Pathway report 3.7 p32
The Integration Architecture shall allow the automated integration of legacy and non-conformant data into the system.	Interfaces	Pathway report 3.7 p33
The Integration Architecture shall allow the NDT catalogue to be continuously refreshed on a ‘just in time’ or periodic basis as required for the data type.	Architectural Principles	Pathway report 3.7 p33

Requirement	Categorisation	Source
The Integration Architecture shall provide continuous, automated testing of compliance of the published resources within the framework.	Data	Pathway report 3.7 p33
The Integration Architecture shall allow access to the data remotely or as a local working copy depending on access permissions.	Data /Security	Stakeholder interview
There shall not be a central authorisation mechanism, but it shall be distributed as required in the ecosystem.	Security	Stakeholder Interview
The Integration Architecture shall be architected and engineered such that it provides a low barrier to entry for provider organisations.	Architectural Principles	Stakeholder Interview
The Integration Architecture shall use open standards where they fulfil the requirement and look to develop open standards where they are not.	Standards	Stakeholder Interview
The Integration Architecture shall support the monitoring of the quality of the data that is part of the National Digital Twin as well as the data itself	Data / Security	Stakeholder Interview
The integration architecture shall support assuring the quality of data proposed to be included in the National Digital Twin before allowing it to be included, including that the quality of the data is specified and that it conforms to the Foundation Data Model and Reference Data Library.		Stakeholder Interview

Table 2. High Level Requirements

Appendix B
Data integration patterns

There are many different ways that integration at the data level can be achieved when exchanging and sharing information. The management of the schemas and the identification of overlapping data scopes is key to information flows that provide the mechanism to successful integration. Figure 11 shows the basic concepts to a methodology developed to illustrate these concepts.

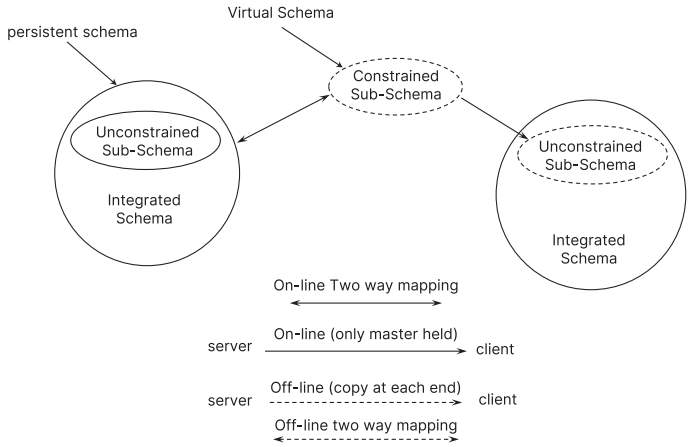


Figure 11: Data Integration Approach – Basic elements

It shows how information is exchanged and shared between repositories that may have persistent schemas or not. It also shows the schemas for the exchange mechanisms (may be messages, streams, or interface calls). For example, a number of systems that share a database is shown in Figure 11. Each of the Applications or systems (labelled 1, 2 and 3) that share the data in the Shared Database, each integrate into their own schema a subset of the overall data in the shared database, labelled respectively X, Y and Z. The connections show that the integration is online and synchronised outwards only from the Shared Database to 1, 2, and 3.

The application of this methodology to the recommended architecture pattern defined in this document is shown in Figure 13 below.

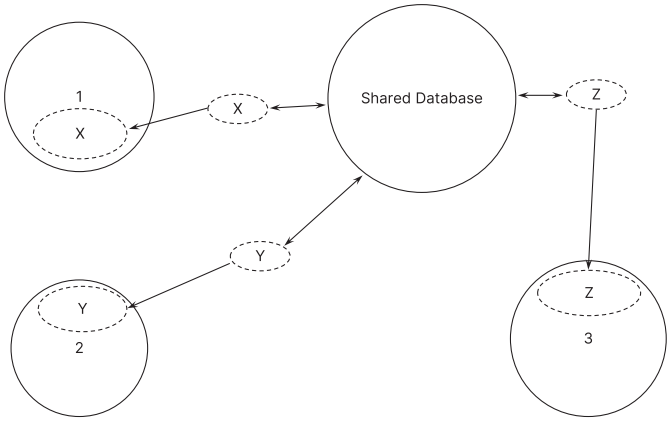


Figure 12: Shared Database

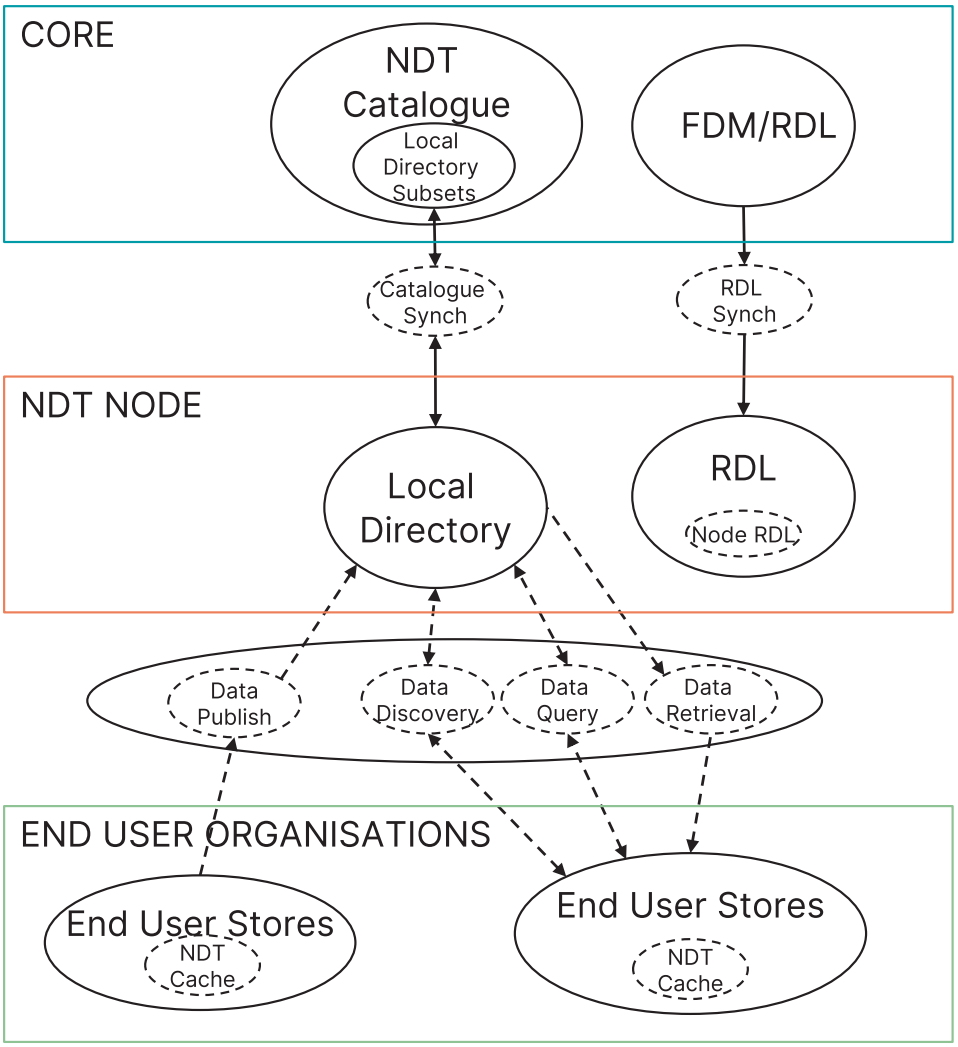


Figure 13: Data Integration Pattern for the IMF Recommended Architecture

Footnotes

1. https://en.wikipedia.org/wiki/Crypto-shredding
2. https://en.wikipedia.org/wiki/Decentralized_application – note this is very early stage technology.
3. https://en.wikipedia.org/wiki/Record_linkage
4. Enterprise Data Header (dni.gov)
5. https://www.ncsc.gov.uk/blog-post/zero-trust-architecture-design-principles
6. https://en.wikipedia.org/wiki/Universally_unique_identifier
7. Explanation given here https://blog.identityautomation.com/rbac-vs-abac-access-control-models-iam-explained
8. https://en.wikipedia.org/wiki/Transport_Layer_Security#TLS_1.3
9. Key Management Service – Amazon Web Services (AWS)
10. Azure Key Vault Overview – Azure Key Vault | Microsoft Docs
11. https://en.wikipedia.org/wiki/Crypto-shredding
12. https://www.cdbb.cam.ac.uk/news/blog-bs-en-iso-19650-52020-supporting-secure-future-digital-construction
13. https://en.wikipedia.org/wiki/OAuth#OAuth_2.0
14. https://en.wikipedia.org/wiki/OpenID
15. Security Assertion Mark-up Language – Wikipedia

References

Hetherington, J., & West, M. (2020). The pathway towards an Information Management Framework — A 'Commons' for Digital Built Britain. doi.org/10.17863/CAM.52659

Cook, A. (2020). Implementation of fine-grained information management controls — Risk management in a distributed, integrated information environment. [Unpublished manuscript]

ISO. (2018). Organization and digitization of information about buildings and civil engineering works, including building information modelling (BIM) — Information management using building information modelling — Part 1: Concepts and principles (ISO 19650-1:2018). https://www.iso.org/standard/68078.html

Acknowledgements

Authors

John Kendall

Contributors

Karen Alford

Ian Bailey

David Bell

Alastair Cook

Mark Enzer

Jim Geatches

Ian Gordon

Steve Kochli

Miranda Sharp

Matthew West

cdbb
Centre for Digital Built Britain

The National Digital Twin programme is funded by the University of Cambridge and the Department for Business, Energy and Industrial Strategy via InnovateUK, part of UK Research and Innovation. This research was enabled by the Construction Innovation Hub(the Hub). The Hub is funded by UK Research and Innovation (UKRI) through the Industrial Strategy Challenge Fund (ISCF).

Kendall, J. (2021). National Digital Twin: Integration Architecture Pattern and Principles. doi.org/10.17863/CAM.68207

CONSTRUCTION INNOVATION HUB

Department for Business, Energy & Industrial Strategy

INDUSTRIAL STRATEGY

UK Research and Innovation