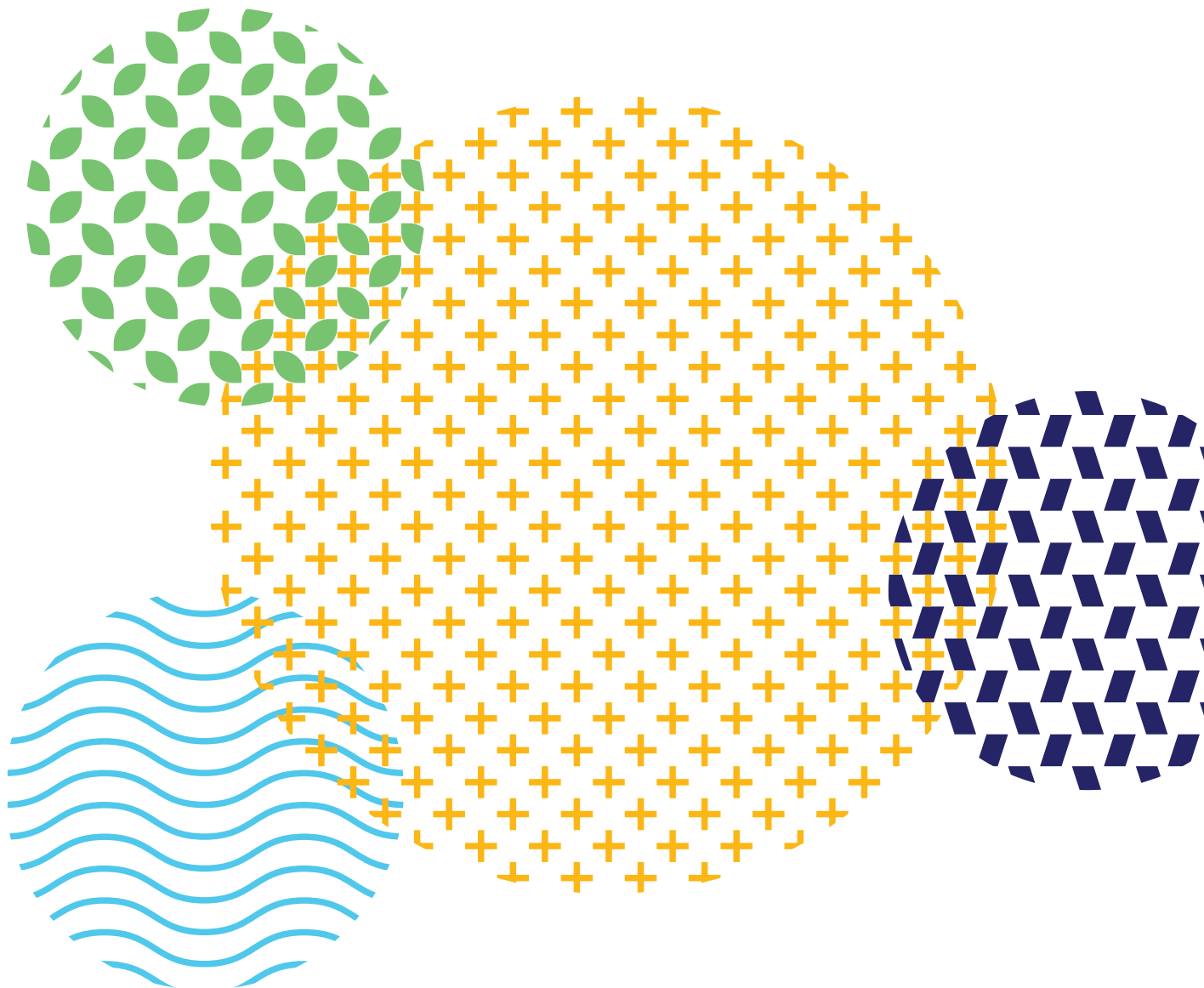


The pathway towards an Information Management Framework

A 'Commons' for Digital Built Britain



Foreword



Dame Wendy Hall FRS FREng
Regius Professor of Computer Science
University of Southampton

In its seminal publication, “Data for the Public Good”, the National Infrastructure Commission set out a vision and challenge of a National Digital Twin: a Digital Twin of National Infrastructure, to provide quality information supporting improved decision taking by those developing, operating, maintaining and using infrastructure and the services they provide to citizens.

Quality information means that it is fit for purpose for the decisions it supports. While properties like accuracy, timeliness, completeness, and provenance are determined when the data is created, properties like availability, clarity, and consistency need to be determined in advance: they don’t happen by accident, only on purpose.

Clarity means that when we see a piece of data, the meaning is unambiguous, and consistency means that we use in common terms and definitions for the same things, and have in common one way of saying something so we do not have to translate data we receive before we can use it. For a National Digital Twin this has particular challenges as we will need to achieve this clarity and consistency across multiple systems in multiple organizations and multiple viewpoints, meeting the requirements to support all of them.

Availability means we have the technical capability to get data from where it is created to where it is needed, being mindful of security so that sensitive data is only available to authorised users, but public data is available to all. We won’t be providing the infrastructure, but we do need to specify the architecture and protocols that allow consistent implementations, in much the same way as the World Wide Web is defined.

To address these issues, “Data for the public Good” recommended the setting up of a Digital Framework Task Group to establish the Digital Framework that would provide the technical basis to achieve this. This report is a critical milestone that sets out what the Digital Framework is, and the steps that need to be taken to put it in place. This means we are now ready to start the development of the Digital Framework as a critical step towards a National Digital Twin. However, even when we have a Digital Framework, it is not the same as having a National Digital Twin, so there is a Roadmap for how to move towards a National Digital Twin, which we need to implement, of which developing the Digital Framework is a part.

These are exciting times as we take our first steps towards a National Digital Twin where information is available, reliable and accessible.

Contents

Foreword	3
1. Executive summary	4
2. Introduction	6
3. End state: what the future holds	10
4. Pathway to change: how we get there	34
Initial tasks to establish the Commons	42
References	56
Acknowledgements	57

1. Executive summary

This report is directed at a technical audience who are interested in developing The Gemini Principles to deliver public good and commercial benefits arising from a National Digital Twin. This report is published alongside a report for a more general audience.

The value of information which is managed effectively and shared securely so that the people can make the best decisions is becoming increasingly well understood. The Gemini Principles set out the guiding values for the creation of a (national) system for connecting digital assets but are designed to enable competition on delivery, encouraging innovation and development over time. That system, the Information Management Framework, and a proposed approach to its creation is discussed in this paper which has been created by a group of information management experts working with the National Digital Twin programme.

We have agreed that an appropriately functioning framework which allows digital twins to connect, would comprise;

Foundation Data Model (FDM)

a consistent, clear understanding of what constitutes the world of digital twins

This ontological model should be able to describe general concepts independent of a problem domain and as a result the work of the NDT programme includes working with twin builders and domain experts to explore and propose structures of relationships to be held within and between digital twins, models, datasets, and physical twins.

Reference Data Library (RDL)

the particular set of classes and the properties we will want to use to describe our digital twins

The specifics of particular controlled vocabularies (taxonomies) and how common words are used, providing standardised choices for differences arising from sectoral and disciplinary legacies.

Integration Architecture (IA)

the protocols that will enable the managed sharing of data

these protocols will include components such as

- A discovery protocol - Allowing for efficient retrieval of digital twins across distributed providers.
- An authorisation layer - Implementing a security model.
- A data transformation and validation engine - partially automating the integration of incompatible data and providing compliance testing.

There is broad agreement on the components of the framework, but different methods of adoption are proposed. What is clear is that the benefits to the economy, society, business and the environment from enabling the connection of twins together at a national level are significant. The programme is focused on learning from those who are already active in this space and sharing best practice both across domains and internationally to progress.

We will work with all of infrastructure, new and long-standing assets, data that is from recently deployed sensors and legacy information to deliver the platform for human flourishing envisioned in the NIC report.

What is the “Commons”?

A NDT requires information to be compatible across the built and natural environment. This will need us to move beyond data exchange to data integration. It will also demand interventions to enable curation and mapping of existing and future models and data.

This will require major effort to create a shared national resource, essential to enabling a secure NDT but largely invisible to the user community. This resource is the technical core of the Information Management Framework, we call it the “Commons”.

2. Introduction

In this report, we identify the requirements for the technical core of an Information Management Framework that supports secure and resilient exchange, inter-operability, and integration and linking of data and models across the built and natural environments at a national level.

We then go on to identify a series of work packages to deliver this framework. Centre for Digital Built Britain (CDBB) has developed a prioritised plan for the delivery of an information management framework for the built environment, and increased capacity and capability within the UK to lead the development of digital twins, and connected digital twins. The roadmap is supported by leaders from the built environment chairing working groups to support the five work streams of the roadmap, approach, commons, governance, enablers and change. This report forms part of deliverable 1.8 from the National Digital Twin programme roadmap. The roadmap is the work plan of the National Digital Twin programme (NDT) and follows the publication of the Gemini Principles [1], which sought to build a consensus on foundational definitions and guiding values for information management across the built environment.

This report is aimed at experts in the domains of information management, mathematical and computational modelling of complex systems, the built and natural environment sectors and national, regional and local infrastructure, and it is supplemented by a document for a more general audience.

This document was created and managed as a working document. This enabled the authors, acting on behalf of the National Digital Twin programme, to identify and manage a number of perceived points of controversy within the scope of the information management framework. Expert workshops were held to discuss those points, and this document is now the culmination of those workshops and the work done on this document between them.

Digital twins are digital representations of something physical or intangible – such as a system, process or object (physical twin). They enable better outcomes through better-informed decisions at all stages of the life cycle. The benefits extend from physical asset life-cycle management to user experience, more informed policy decisions, better responses to new technology and regulation and the facilitation of new business models. A digital twin is more than just a model of a physical asset; it provides context (i.e. the relationship between the asset and its environment), connectivity between digital and physical assets (in at least one direction) and the ability to monitor the physical system in a timely manner.

Data flows between the physical and the digital twin enable insights and create the opportunity for positive interventions within the physical twin.

Digital twins have been around for years (though not always with that name) delivering benefit in sectors from manufacturing to Formula 1. Now that they are gaining traction in the built and natural environment, the potential is emerging to connect digital twins together to make cross-sector observations, decisions and interventions, and to evaluate the impact of those interventions. Enabling the secure, resilient and reliable integration of twins in different domains, allowing the sharing of information, is the goal of the National Digital Twin project. It is in this way that we will facilitate the creation of an ecosystem of loosely coupled digital twins, discoverable through the protocols defined, brought together at need and queried, which together can be seen as a composite National Digital Twin programme.

Much additional work will be required by other themes within the wider NDT programme. In addition to delivering on the technical objectives described in this document, we will need to develop an understanding of how the modelled assets contribute to the benefits they enable. Security is a central pillar of the roadmap for delivery of the National Digital Twin programme, and it must be considered throughout.

While this document focuses on the technical challenge, digital twins will not deliver value to society without a sophisticated understanding of the social systems with which they are interdependent. An effective Digital Built Britain must not be solely focused on technical and physical aspects, but it will also need to carefully consider the societal impacts.

Information management and governance policies and standards will also be needed to support integrated asset management and use thereof, to maintain coherence between the digital and physical elements of a twin.

There will also be a skills and training demand to realise this vision: many organisations, perhaps most, will be utilising legacy approaches. The difficulty of mapping legacy approaches to these standards may be significant.

2.1. The vision of a Digital Built Britain

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 [2] a National Digital Twin could provide insights that enable 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 potential effects of changes to our physical environment before disruptive interventions are made, as will linking between the legacy approaches of different organisations.

The NDT will not be a large, single model of the entire built and natural environment. Instead, it will consist of multiple consistent digital twins integrated via securely and appropriately shared data, using protocols and formalisms to solve questions in a distributed way. The NDT will enable infrastructure professionals to make better decisions at project, asset, network, system enterprise or place level, and supporting the public in their use of infrastructure systems, and the built and natural environment.

An NDT therefore requires information to be compatible across the built and natural environment, presented in consistent formats to allow for sharing and integration between different digital twins. This will need us to move beyond data exchange to data sharing and integration. It will also demand interventions to enable curation and mapping of existing and future models and data. However, beneath the surface, it will require major work to make it a reality. This will need up-front work to create the shared understanding, reference data and common data models and protocols that those contributing to the NDT ecosystem will need. These elements are referred to as the "information management Commons" (hereafter also referred to as "the Commons"): essential to the development of the NDT but largely invisible to the user community. What we are proposing is an infrastructure to facilitate better integration of digital twins, not a shortcut, and it will be vital that data providers and researchers continue to consider data security and risks as they share and integrate data sets.

This work will promote the evolution of a market in services related to digital twins that will enable more effective ways of managing models and data for the public good. This will help to shift our culture towards treating information as a valuable asset, and support stakeholders in managing it properly to ensure it is fit for purpose.

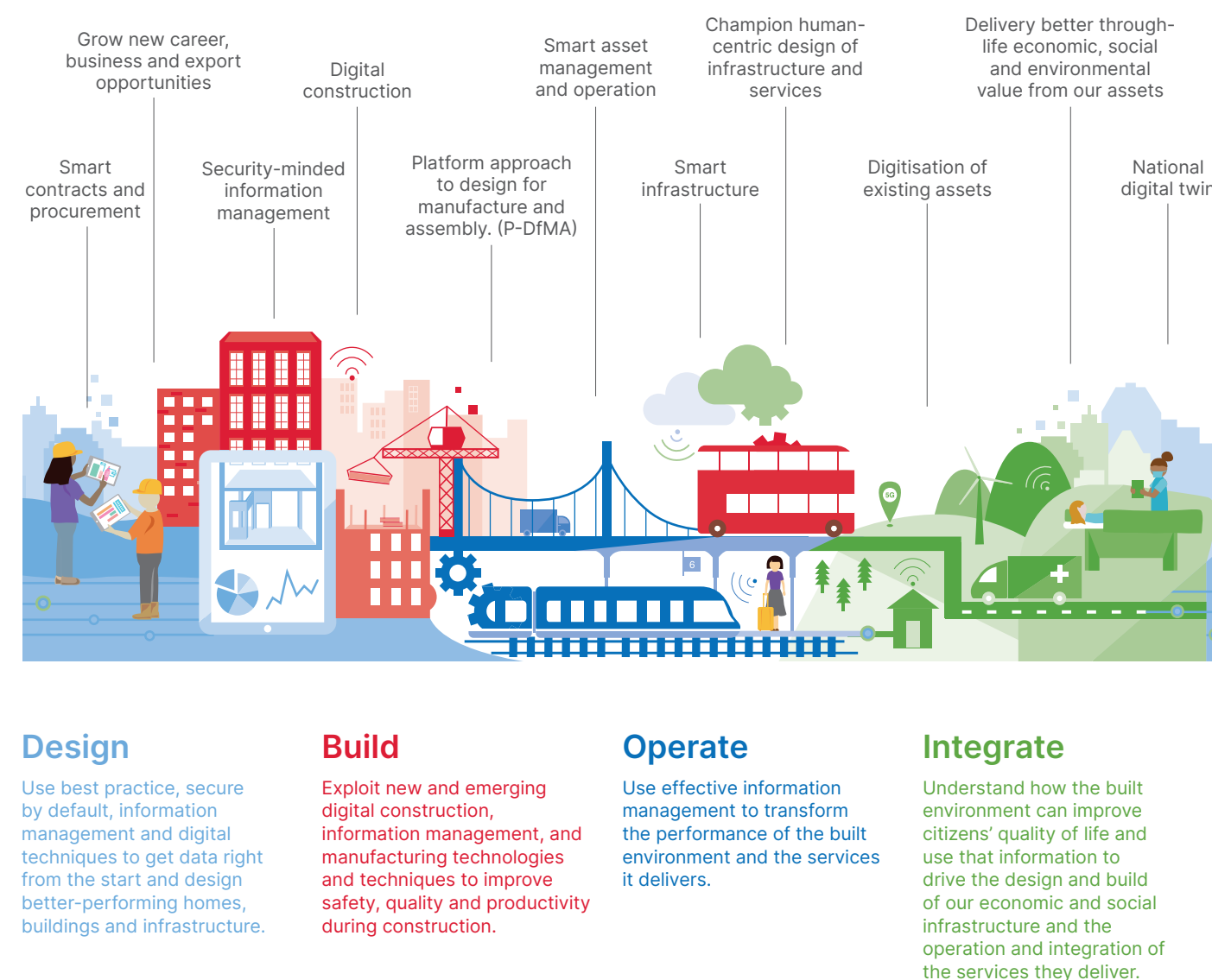


Figure 1: Vision for Digital Built Britain

3. End state: what the future holds

We now set out a number of examples: these are not exhaustive, but they are illustrative of the kind of outcomes an NDT could deliver, through facilitating data sharing for more effective modelling. These are deliberately cross-sector, address examples in both rural and urban contexts, and address infrastructure, both static and in-use.


These are provided to show the long-term potential of an NDT. In section four, we will set out a delivery programme that will create the enablers that will allow emergent markets to also provide quick wins.

We begin by considering how a distributed NDT would allow us to query across multiple data sources. Consider, for example, how we could obtain a timely and accurate answer to the query, “Which tower blocks in the UK have the types of cladding identified in the Hackitt inquiry as a high fire risk?” [3].

This would require information about which buildings were tower blocks and the type of cladding each tower block has, to be obtained through database-like queries.

We note, however, that this does not require the data to be held in a single centralised asset-database. The premise of the NDT is that by developing a common data model, reference data, appropriate protocols and shared data formalisms, the question could be solved in a distributed way. To be able to query the information effectively, and to support its use by multiple parties, it needs to be held as, or to be able to be transformed into, consistently structured, classified machine-interpretable data so that multiple users can query it simultaneously. Models to query the NDT can be plugged into its infrastructure, extending their capability.

We must also be able to easily discover the existence of multiple sources, query them through online services, and be informed of their provenance. If this data were accessible to authorised users, with appropriate security protocols and oversight, it would be possible to write a distributed query across all local planning databases holding such information to find out which buildings, and in particular, which tower blocks, had used the same or a similar type of cladding. Such a query will only work if all the data can be addressed in the same way – by understanding how to extract, transform and integrate varied data into a common data model and reference data.



We must also be able to easily discover the existence of multiple sources, query them through online services, and be informed of their provenance.

Example 1:**3.1. Complex decision support for retrofit of at-risk buildings**

Digital twins allow for more complex usage patterns than just queries.

Continuing our Hackitt inquiry example, we could imagine that, having run a query to find all the tower blocks with the given cladding, we also then ask, "OK, rank each building by its fire risk". To resolve this query, an automated system would have been able to:

- Define what we mean by a tower block (i.e. the characteristics that make it a tower block rather than another type of building).
- Retrieve structural, compartmentalisation and architectural models of each building, and any existing data on fire risk.
- Transform these into dynamic finite element models allowing for fire propagation analysis.
- Consider how socio-economic models of occupancy drive fire risk parameters: for example, how ignition risk is influenced by maintenance of household electrical.
- Retrieve information on occupancy levels over time for each building from building instrumentation and telemetry.

- Combine this with adapted agency models for human behaviour, to provide guidance for individuals on evacuation in the case of an emergency.
- Retrieve transport models for the relevant cities allowing for travel time prediction for emergency vehicles.
- Understand the provenance and uncertainty in all the above data "ingredients", to inform an evaluation of the reliability of the conclusions being drawn.
- Compose these into a risk model, enabling automated decision support for prioritisation of retrofit expenditure.

Where these information sets are accessible in a consistent format, this combination will be simple. Elsewhere, a wide variety of tools and methods will be required to collate and prepare the data. We also note that many of the ingredients needed to solve this problem will also be of use in understanding and managing other risks: thus, as the ecosystem of digital twins grows, this will drive efficiencies in that system.

Example 2:**3.2. Regional resilience, response and simulation**

How could a composite, integrated digital twin of a local area improve the resilience of the wider region?

Bridges and other assets make up key elements of national infrastructure, but some are more vital than others. The impact of a single critical asset failure can have far-reaching consequences beyond just the immediate area.

In a set of floods that struck Cumbria in 2015, the Cockermouth and Tadcaster bridges were taken out of action when their structures became unstable, having been undermined by the scour action of floodwater eroding abutments. The bridge failures impacted movement, utilities, communications and power services over a huge area, hindering relief efforts and forcing first responders to take long diversions to cross the valley (Figure 2).

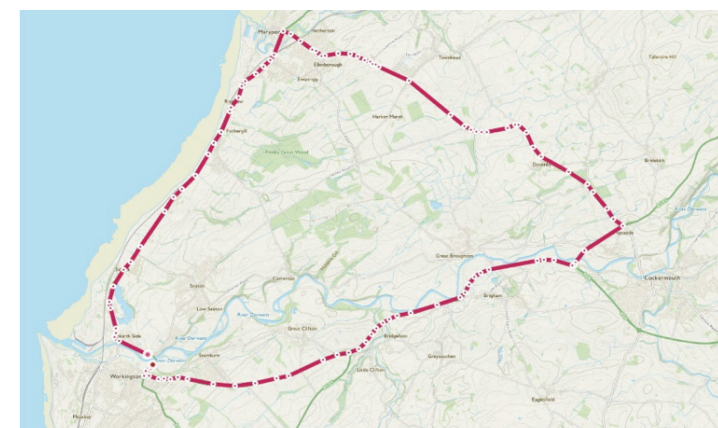


Figure 2: Diversion following bridge failure in Cockermouth 2015

The information necessary to flag these bridges as flood risks had existed at the time, but it was held in disparate data sets across different organisations, hindering pre-emptive action.

What if we could map out the interdependencies of different systems, building a composite model to test how critical events could impact infrastructure and the wider area, monitoring preparedness and prioritising efforts to shore up vital structures?

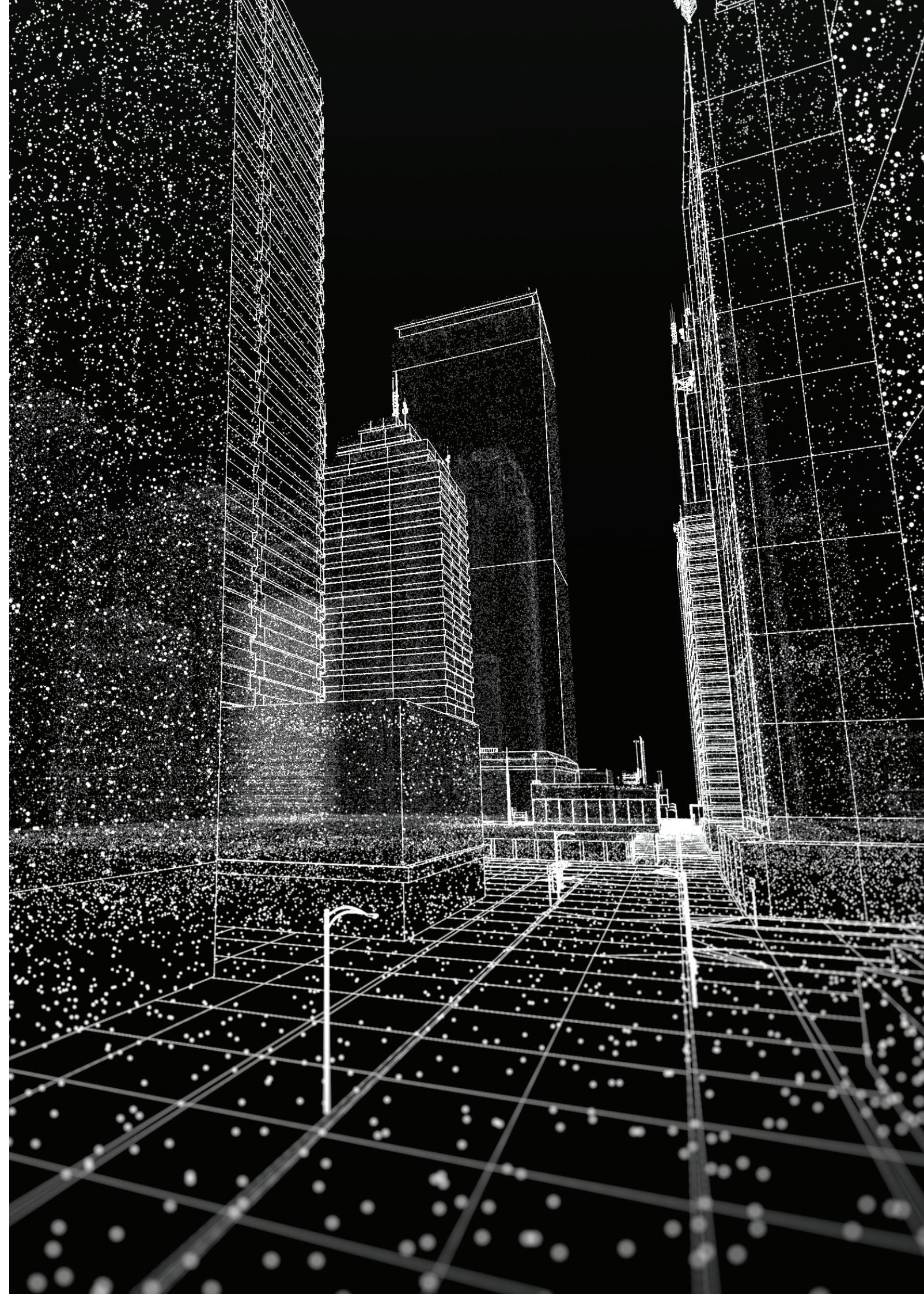
By simulating disasters and major events, we could build more targeted guidance for first responders, and build in tools to allow for early detection. Furthermore, what if we could, following such interventions, use a digital twin to measure the impact of the interventions and justify the investment.

Achieving something on this scale would require a model that can:

- Integrate transport data (both fixed infrastructure and usage data) with communications, utilities and power, to map out regional infrastructure.
- Join it to socio-economic data to model where and how the impact of failure of key infrastructure assets would be felt.
- Use environmental and meteorological models to indicate where critical events and weather conditions could have the potential to cause significant disruption.

- Simulate events to spot how they can be detected early, and inform detailed response plans.
- Understand the provenance and uncertainty in all the above data “ingredients”, to inform an evaluation of the reliability of the conclusions being drawn.
- Operate in a security-minded manner, interpreting and enforcing security requirements so as to limit access to data and any subsequent analysis to those that have a legitimate need and thus limit the model’s use as a hostile reconnaissance tool that could be used to target vulnerabilities and/or maximise the harm from hostile or malicious acts.
- Monitor the bridges, roads, pipes and wires that the above analysis reveals to be critical with a sensor network linked to live structural models, to understand their stresses and risk of failure.
- Loop feedback on interventions back into the model, iteratively improving it over time.

By simulating disasters and major events, we could build more targeted guidance for first responders, and build in tools to allow for early detection.



Example 3:**3.3. The citizen-centric data model**

People are missing and/or misrepresented through the assumptions made about models used to guide investment into infrastructure.

The upgrading of critical infrastructure requires the interpretation of data that represents the living conditions of citizens. As building automation systems analytics are increasingly used to represent the living conditions, location and behaviour of citizens, people and places can often find themselves unaccounted for in key local and national data sets, missed out of service improvement plans and government initiatives. It can be difficult to spot where, and how, people are being left behind. To tackle this, we need to build a model that combines different analytics tools to understand how the whole built environment affects citizens, identify where citizens/asset users/asset owners are missing in the data, and to better understand the impact of interventions.

For example, local authorities and utility providers hold data on the distribution and quality of services in their area, as well as on outcomes for the communities they represent. For example, if I build a new bridge, what would the change in transport access be, and optimisation questions?

For example, where would the optimal place for a new childrens' hospital be? We could consider data such as: access to transport, water / air quality, mobile / internet connectivity, crime, health and employment.

To bring real benefit, our digital twin would need to:

- Access and combine infrastructure/ utilities/communications/transport data with environmental data sets and models of citizen outcomes.
- Address the missing data imputation problem in national citizen models.
- Understand the representation of citizens and communities in local/ national data sets, to identify where different groups may have been left behind.
- Explore trade-offs arising from modelled interventions and provide evidence for discussion and choice of potential improvements to citizens' outcomes.
- Recommend interventions based on the potential to improve citizen outcomes based on evidence.
- Measure the impacts of interventions over time, feeding back into the digital twin model.

This is a citizen-centric data model, with appropriate measures employed to protect citizens from unauthorised or illegal targeting methods, taking into account commercial, security and privacy sensitivities of the sharing of data relating to citizens' use of assets and services. Much of this data should be available from service and infrastructure providers, and regulators can push markets to be more open with their data. Where appropriate, the NDT should act to promote commercial practice for the public good, in line with the Gemini Principles [1].

We need to build a model that combines different analytics tools to understand how the whole built environment affects citizens, identify where citizens / asset users / asset owners are missing in the data, and to better understand the impact of interventions.

3.4. The scale of the task

Much of the data and technology to enable digital twins to solve such problems already exists (for example, AI can already be used to predict flash floods [4]), but a number of challenges need to be overcome to allow these benefits to be realised, including inconsistencies in the quality of the data.

Beyond the challenge of sharing and integrating them, models of this size would require considerable resources to develop, create and operate. It would be both wasteful and unrealistic to imagine this could be delivered in a single focused project. Not only would working in isolation incur significant design, implementation and maintenance inefficiency, but such a programme would also, in aggregate, be on a scale comparable to a moon landing or new particle accelerator.

There are lessons to be learned from the success of the development of the world wide web - the delivery of a global-scale solution through massive, distributed activity, made coherent through standards for interoperability and information exchange: a “web of digital twins” to complement the Internet of Things (and with similar questions around misinformation).

Achieving this will require:

- Interoperable models and standards for structuring data across the built environment, and the ability either to automatically extract and transform data into such a consistent form, or to update local data sources into consistent formats.
- A data model that extends beyond the purely physical/functional representation to address attributes related to the legal, regulatory, societal, ethical and temporal contexts within which the assets operate or may be required to operate [5].
- Integrated asset and configuration management to manage the spatial and temporal (and other) changes over the physical asset's life cycle.
- Protocols for abstracting services, including those that provide real-time streaming data, and also those services that allow us to execute the models and resolve the queries needed to be able to ask questions of digital twins.
- Protocols for managing the distribution of outputs from queries or model resolution such that they can be used as inputs in further queries in consistent and automated (or, where that is not achievable, semi-automated) ways.
- Approaches for deploying digital twins onto computational infrastructure, perhaps cloud-based, upon demand, and scaling capability when required.



- Protocols for managing, securing and controlling access to information: which actors, in which roles, can access which data sets and invoke which models?
- Information management and governance policies and processes that support integrated asset and configuration management so as to maintain coherence between the digital and physical elements of a twin and enabling integration of cross-domain twins.

These protocols and data models form the NDT “Commons”. It is the distributed ecosystem of digital twins, discoverable through the protocols defined, brought together at need and queried, that forms the NDT itself.

A principled approach can be taken now, to start delivering key components that will be an essential part of this future – these components will themselves provide value along the way through application to support simpler challenges, linking twins and models, gradually increasing in sophistication and complexity. However, delivery of the capability to specify the queries and pipelines that will compose such a complex composite twin from published components is an enormous task, requiring decades for it to reach fruition, and continuous improvement and maintenance as further data becomes available and technology advances.

It should be noted that the development of an NDT requires coupling of models between organisations and sectors, such as between buildings, utilities, infrastructure, transport, social and environmental models. It is also noted that an NDT must accommodate the human and natural, as well as the built, environments (with appropriate consideration to personally identifiable information, and its protection under the GDPR). Complex supply chains will need to exchange data to ensure that sufficient data is available to those who need it. Such intricate cross-organisation couplings show the requirement for, and expected value to be gained from, facilitating information interchange and integration between digital twins (and the organisations that contribute to and benefit from them). It is worth noting that the use of such data may come with strict caveats or specific handling restrictions, as a result of the presence of personal, commercial or other security-sensitive information. Any such requirements have to be agreed in advance and handled in such a way that all parties involved maintain mutual trust.

What kind of things do we want to be able to say about or ask of our digital twins? We will need to query with tools that can align the rigour required for machine reasoning to the flexibility of human language. Defining a simple database table for every kind of entity, from bricks to bridges, from mathematical models to compute clouds, from railway lines to regulations, or from owners to citizens, will be impossible.

In particular, to build this federated system of digital twins, we will need:

A Foundation Data Model (FDM):

a consistent, clear understanding of what constitutes the world of digital twins, and how we want to be able to formally describe them and their applications in a machine-interpretable way.

A Reference Data Library (RDL):

the particular common set of classes and the properties we will want to use to describe our digital twins. In practice, this could take form through a federation of domain-specific libraries, with a single common protocol. Many of these definitions and properties already exist, and in some cases we can use automated tools, such as semantic data mining, to help to collate them (as described in Task 3). Reference libraries do not remove the need for subject-matter experts to interpret the results and extract policy insights or commercial decisions.

Furthermore, multidisciplinary skill sets must be fostered to interpret results that span the boundaries of traditional disciplines.

An Integration Architecture (IA):

the protocols that will enable the managed sharing of data, the production of models, the scripting of queries and the analysis, interpretation and application of model outputs. This will include the protocols that catalogue digital twins and make them discoverable, and the minimum information sets required by those offering digital twin services.

Whilst the FDM will provide the structure for “sentences” (i.e. describing the relationships between things whether physical or intangible), the RDL must provide the controlled vocabulary (i.e. the dictionary of terms with defined meanings/uses) that allows us to compose the sentences about the things. We can then define the world using simple sentences such as <beam 72> <is part of> <the British Library> or <beam 72> <has elastic modulus> <200 GPa>. The RDL ensures that different parties understand a <beam> as the same kind of entity – not, in this case, a laser beam.¹

We call the technologies that allow us to formalise such statements, making unique machine identifiers for relations such as <is part of>, “semantic modelling tools”. The grammar expressing how these terms may validly be used, and the relationships within it are called “ontologies”. Our ontology will combine the use of the FDM and sets of reference data from the RDL to enable us to create sentences or statements about things in one or more domains. A number of standards already exist for these.

The following sections describe each of the FDM, RDL and IA in turn.

¹ The reader should be aware here that the above example is illustrative, and more careful work will have to be carried out to develop real semantic methods.



3.5. A Foundation Data Model: clearing up the concepts

Our Foundation Data Model will need to address the questions proper to an upper ontology, which can describe general concepts independent of a problem domain. Our FDM should be able to provide answers to:

- **Time, space and place:** How does the ontology deal with time and space-time? How does the ontology deal with places, locations, shape, holes and a vacuum?
- **Actuality and possibility:** How does the ontology deal with what could happen or what could be the case, such as where multiple data sets give conflicting stories on the behaviour of a network?
- **Classes and types:** How does the ontology deal with issues of classification?
- **Time and change:** How does the ontology deal with time and change?
- **Parts, wholes, unity and boundaries:** How does the ontology deal with relations of parthood?
- **Scale and granularity:** How does the ontology deal with scale, resolution and granularity?
- **Qualities and other attributes:** How does the ontology deal with qualities and other qualitative attributes?
- **Quantities and mathematical entities:** How does the ontology deal with quantitative data and with mathematical data and theories?
- **Processes and events:** How does the ontology deal with processes?
- **Constitution:** How does the ontology deal with the relation – sometimes referred to as a relation of “constitution” – between material entities and the material of which, at any given time, they are made?
- **Causality:** How does the ontology deal with causality?
- **Information and reference:** How does the ontology deal with information entities?
- **Artefacts and socially constructed entities:** How does the ontology deal with artefacts (e.g. engineered items) and socially constructed items like money and laws?

To answer these questions, we will need to explore the implication of these questions for our domain, proposing that the structures of relationships be held within and between digital twins, models, data sets, and physical twins. This document does not attempt to answer these now: examining and mining existing digital twins will be the work of the project itself.

- A digital twin is more than just a collection of pieces of data that describes the world. How do we describe the relationship between a digital twin and the corresponding elementary pieces of data?
- How do we describe the domain of validity of a digital twin, including any assumptions or simplifications of real-world behaviour? This may include describing assumptions about the underpinning physics, engineering, biology or sociology that influence the way that the asset operates. How do we define the boundaries of validity of models to ensure that models are used and composed only in ways that are meaningful? Especially for “black box” models, how do we ensure the validity of inputs and outputs and mitigate the risk/damage of erroneous outputs, especially when users will not be close to the development decisions made in the creation of the models.
- How will the data models used by an existing digital twin be validated and interpreted so that mappings and transformations can be established prior to seeking to integrate the twins?
- What is the relationship between a twin and the physical, mathematical or computational model(s) that underpins it? How do we define the non-physical parameters describing the mathematical model used, such as a grid resolution?
- What is the relationship between a digital twin, and the kind of things it describes? Is this a twin of the make and model of my car, or of my specific car? If the latter, what do we call the kind of thing that is a potential twin of all such cars, before I connect it to the telemetry from my specific car? To what extent does such telemetry actually need to be real-time or would it be sufficient to periodically collect, to analyse performance and arrange maintenance? How do we aggregate models that operate at very different tempos – from seconds, to hours to days?
- How do we make statements about time? How do I talk about discrete time periods like “on Thursdays” or “in Summer” or “FY 20/21”? How do we describe and capture change over time? How do we model future operations to assess the impact of planned changes, without disturbing the current operating model?

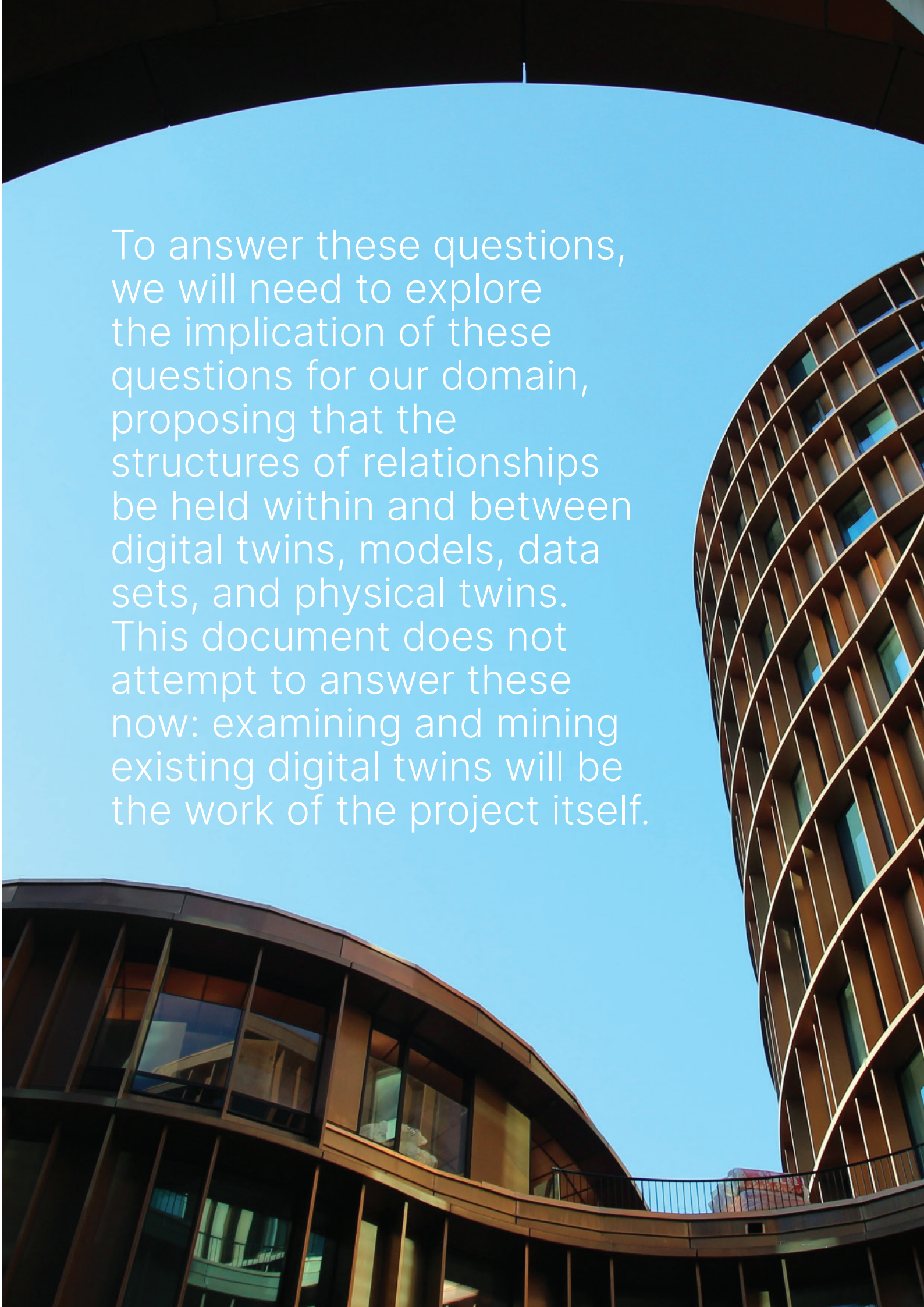
- How do we break down the physical world into parts? What is the relationship between a twin of a component of a city, such as a building, and a twin of a city? How do we reference the “coarse-graining” that takes place when we have different models, at different resolutions, that overlap in the aspects of the physical world that they describe? How do we aggregate models, particularly in circumstances where there may be modelling gaps? How do we deal with missing information, or unconnected assets?
- How do we handle uncertainty? How should we best manage the difference between “measurement uncertainty”, “variability within a class”, “variation over time”, “environmental noise” and so on?
- Some models will be mechanistic, based on known understanding of the physical world. Others will be purely empirical: based on maximising goodness-of-fit from models to information without incorporation of domain knowledge. Many will be a mix of these paradigms [6]. How will this aspect of the use of digital twins be reflected in the ontology? How reliable is each paradigm, and how can a lack of reliability be taken into account?

It is also worth noting that the scope of the data to be described covers more than just the digital twins themselves. Our data requirements also include the following:

- How do we handle versioning? Should version histories be curated indefinitely? How do we handle archiving and ultimately removal of out-of-date data?
- How is invalid data corrected? What audit trail is needed and how will data

ownership be managed and maintained? Are statements recorded together with the identity of the person or organisation making the claim, so that the provenance of information is tracked and unreliable information can be managed? How do we correctly handle missing, invalid or inaccurate data? Outlying or erroneous data can sometimes still be useful when analysed in a different way.

- What relationships do twins have to their authors? Who owns them? Who can know what about them? What kind of roles and actors are there?
- How do we describe not just the models themselves, but the methods that are used to derive insight from them? Models live alongside visualisations, interfaces, deployments and so on, which also need to be described and exchanged.
- How do we describe the uses to which models have been put? Will we need to log each question asked of a digital twin? This will facilitate audit and meta-analysis, and save on computational time lost when re-running old studies, but may have information governance and privacy implications.
- How do we model social concepts related to ownership, rights, legislation and regulation? What are the permitted uses of the model and any licence or usage constraints?



To answer these questions, we will need to explore the implication of these questions for our domain, proposing that the structures of relationships be held within and between digital twins, models, data sets, and physical twins. This document does not attempt to answer these now: examining and mining existing digital twins will be the work of the project itself.

3.6. The Reference Data Library: a vocabulary for describing digital twins

As well as a clear foundational ontology, we also need the specifics of the particular controlled vocabularies (taxonomies) we will use.

While the foundation data model will include core concepts like “space” it will also scale out to several detailed sub-classes (perhaps “room” or “atrium” or “sleeping room”). The controlled vocabularies will also need to address how common words are used.

For example, when referring to a door what elements does this term encompass (e.g. the door leaf, plus door furniture and/or the door frame)? This will provide standardised choices for semantic differences arising from sectoral and disciplinary legacies.

The RDL will also have standards for the minimum information required for its use within the NDT [7]. Our RDL will specify the classifications and characteristics of things. Shared high-quality reference data enables the consistent provision of interoperable data services, to be provided free at the point of use.

It is the RDL that will provide the common reference for parties wanting to exchange data, as it provides a meaningful basis on which to map and interpret each other's data. Consistency is achieved by each RDL being designed subject to the FDM. The RDL, therefore, is not just that part of the RDL published by the information management commons, but rather an

evolving ecosystem of libraries linked together. The resulting ecosystem will form a hierarchical and integrated set of reference data libraries, each building on existing definitions and following the same community-wide information quality management practices.

However, most data, (distinguished from reference data in the RDL) will be provided by third parties, following the protocols specified in the reference data library. Examples include component system or asset- specific information such as the characteristics of a material. These processes and protocols will manage identification of and authorised access to the reference data owned by multiple parties. This data is about particular things, such as where a building is, what the ownership of an asset is, etc. and so on. This changes over time and keeping it up-to-date whilst maintaining its history is expensive. Consistent reference data will encourage organisations to contribute data, and to build businesses around it. There will continue to be different sources of this data, of varying quality and price, including free. This distributed approach will be sustainable over many years, perhaps decades, as technologies and tools change.

The information management commons will provide the description and definitions allowing compliance of published information with the RDL to be assessed. The processes for the use of this information to certify or verify compliance of assets belongs in the governance stream of the NDT programme.



3.7. An Integration Architecture: protocols for integrating digital twins

An Integration Architecture defines the rules and mechanisms to manage the ways digital twins can be combined.

It is important to observe the multiplicities shown on the figure: this is not a single monolithic architecture: there are multiple federated instances of all components. Access to data will be controlled by an authorisation layer that allows data owners to make data visible and available to authorised users.

The final component of our proposed solution will be the digital systems that manage our twins. The NDT Programme will not be building the majority of this reference data architecture; rather, it will be building the protocol to enable an integration architecture such as the one displayed in figure 3 to emerge. Actors will be able to link new and existing twins through this architecture, and query twins and assets throughout the wider ecosystem.

A National Digital Twin enabled by an Information Management Framework

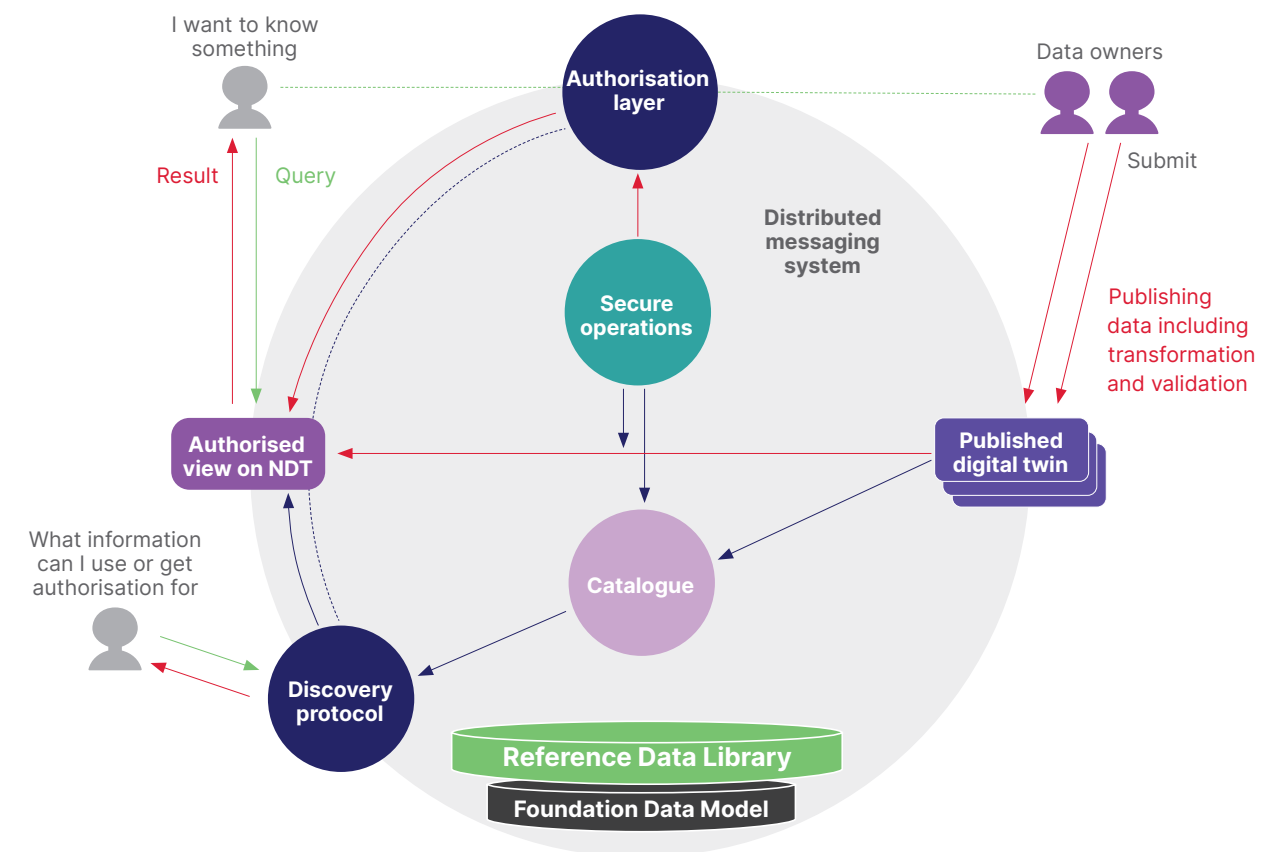


Figure 3: Elements of an integration architecture

Elements will include:

- **An authorisation layer** - Implementing our security model. Consideration will need to be given to how information management, governance and security applies across multiple organisations. For example, an asset owner may be prepared to share different sets of data from a digital twin with a variety of users accessing the twin's data through the authorisation engine, using different integration tools. This limits and governs access to data otherwise discoverable through the discovery protocol. Security concerns may mean that some potentially available models and data sets do not even appear in a particular user's view of the catalogue, since publication of their existence may be a risk to sensitive assets and information. Twin owners will be able to use the languages described in the FDM and RDL to specify the purposes to which their twins or the data produced from them may be put, limiting the types of models that can be conjoined so as to protect intellectual property and commercially sensitive operational data from malicious intent (through both policy and technical measures). This manages communication between digital and physical twins so as to mitigate potential hostile or malicious interference with the assets.
- **A digital twin catalogue** - The comprehensive catalogue of digital twins considered within the National Digital Twin.
- **A discovery protocol** - Allowing for efficient retrieval of digital twins across distributed providers. Providers will need to specify or register with this the information, services and twins they intend to offer. Although the information and metadata itself is distributed, from a user point of view, it will be possible to query this as if it were a single catalogue to allow identification of models and data sets that could be combined/analysed together. Indeed, this should be more than just a distributed catalogue: through the use of data virtualisation users will be able to interact with all the data they are authorised to access as if it were a single database.
- **A query protocol** - Queries on and across digital twins may require significant computational effort. This may require support for query parallelisation and execution in the cloud.
- **A messaging system** - A messaging layer allowing parameters and results to be shared between digital twins, and reference data libraries. This system will act as the link between the distributed digital twins.

Also included in the architecture, although not detailed above, will be two "engines" to maintain the validity and consistency of the model:

- **A data transformation engine** - Not all our data will be available in formats that are compatible with our foundational ontology. We will need to (at least partially) automate the integration of incompatible and legacy data into the system, both once on discovery and "just in time". Permanent automated mappings will allow the digital twin catalogue to be continuously refreshed, while the transformation will be schemed to meet need, with links not needing to be "always on". It will not be possible to develop all of these ourselves: but we will facilitate a distributed solution to this problem by providing tools to support this task (the responsibility to signpost information quality will remain on data providers). In some cases, this will be transformation of real-time telemetry, while in other cases, it will be sufficient to periodically collect and process telemetry to analyse performance and arrange maintenance.
- **A validation engine** - To provide continuous, automated testing of the compliance of published resources with the framework. This must work alongside information providers' own assessment on the quality of their data, set as a "confidence level".

It is important to emphasise that we will be delivering the standards that will allow this ecosystem to function, not the platforms themselves. In order to demonstrate the capability of the Commons, we may need to construct our own examples of transformations of digital twins, of model execution protocols, or of query systems that discover and interrogate digital twins, these; will not be products of the Commons. The reader may find the analogy to the origins of the web useful.

The artefacts forming the Commons are an "HTML" and "HTTP" for digital twins: the first being our reference data library, allowing us to describe twins and reference data; and the second our integration architecture, allowing digital twins to be interrogated and produce results. We may, in passing, also create consumer tools fulfilling the roles of "Netscape" (an early web browser capable of rendering HTML) or tools for twin providers fulfilling the role of "Apache" (a web server capable of servicing HTTP requests) as demonstrators, but these will not be maintained as outputs of the Commons. In the medium term, software and data suppliers will develop their own, standards-compliant, products in these roles.

4. Pathway to change: how we get there

4.1. Our approach

An integrated approach across the programme

The Commons building blocks (FDM, RDL and IA) are just part of the creation and adoption of an integrated information management landscape across the built environment and infrastructural sectors of the UK. Much of the work to share knowledge with, engage and govern the NDT sits in other themes of the NDT programme. The Commons will produce the deliverables that enable the whole programme to establish the processes to evolve and maintain an NDT.

The objective of the NDT programme is not simply to produce a digital framework but to support industry and government in exploiting their information in enabling better decisions.

We will need to build both data-management and governance processes for managing the acceptance of digital twins and reference data as part of the NDT. A distributed, scalable process for verifying that concepts defined in the vocabulary are used correctly, involving both humans and machines, will be required. This will require clear standards, and tutorials, for the use of, and acceptance into, the framework. This documentation will be produced within the Commons stream, though the processes that use it to maintain and regulate the Commons form part of other streams. Beyond the work of the Commons in producing these deliverables, the wider programme will establish processes to deliver that consistency and continuity of service. Although we will need to demonstrate progress by assigning complete-and-finish tasks to deliver the Commons deliverables, it is the communities of practice and the processes that emerge from these that will deliver ongoing value.

Consultative and engaging

We recognise that the development of the Commons has the potential to impact many stakeholders, especially those already active in the area, those with legacy data and systems and those whose businesses currently depend upon such systems. We intend to engage widely and deeply with people with experience and expertise in the creation, curation and use of digital twins and their underpinning data sets. We will also work with those who can see the future potential for using digital twins to access new value. The purpose of such engagement is to maximise value and minimise inconvenience by being consultative and mindful about design decisions made in the creation of the Commons and about the consequences of such decisions. We will work closely with a key group of industry stakeholders throughout the development programme to validate our work. The Commons will benefit from and seek to deliver benefit to any organisation that needs to exchange information with others in ways that they can't do today, using the vision of an NDT as a goal. It will seek to collect requirements from all organisations that willingly participate in Digital Built Britain (DBB), provide opportunity for their direct involvement and be open to all in its progress and deliverables. We will be considerate of impacts on asset owners based on their existing approaches.

Incremental delivery

If we had to identify all the information that will be needed to describe the life cycle and use of the built and natural environment before we start creating the data models and reference data libraries, we would never get started, never mind finish. So, we need an approach that is extensible and tolerant, such that we can start small and expand. To secure continued industry and government buy-in, we must deliver real value as we go.

During development, before release as recommendations for others to build on, our ontologies (FDM and RDL) will be carefully versioned. During this development phase when a non-backwards compatible change is made, we will reload the source digital twins to confirm that the new mappings are valid.

Once released as recommendations for others to use, however, backwards-compatibility breaking changes will impose a heavy cost on users. Therefore, we will take care to test for future backwards compatibility issues against a wide range of scenarios, even ones beyond our expected scope, before we release a new version of either the FDM or RDL. Once components of the Commons are released as external recommendations, we will be change-averse. In particular, with the first release of any component of the reference data library, beyond a small group of collaborating early test adopters who are willing to

accept this risk, any further changes must be fully backwards compatible, extending the model rather than adapting it. This is necessary to achieve the consistency that is the foundation of being able to share and integrate data.

We cannot wait until we have perfected the Commons before allowing it to be used in the real world, but we must avoid imposing change-costs on adopters. It will be necessary to carefully balance these factors, as when we decide to make the first releases, we are recommending that they are fit for others to build against. Data models and information libraries will continue to evolve over time as new technologies develop, demands for processing digital twins expand and refinements are identified. When we are confident that our exemplars establish the work is sound, we will follow national standards development processes and ensure that potentially costly changes to the core concepts are not introduced at a later stage.

Evidence-based

How, then, can we ensure that our models will work when confronted with real data? We must take an evidence-based approach, gathering a corpus of real-world digital twins, and involving asset owners to pilot and test approaches. This will help us to understand and overcome the challenges arising from the use of different legacy approaches.

Building on what already exists

We are not starting from zero: there are already well-known and well-understood ontologies for many of the components of the picture. Members of the programme are keenly aware that concepts of objects, properties and relationships already exist in existing resources and implementations, and that they have complex interconnections. We can build on existing semantic standards from buildings information modelling, urban modelling, computational science, digital rights management and more to synthesise our vocabulary; including national and international standards. However, these will not be mutually consistent, so we may not be able to simply import all of them as-is.

This is where our development of a FDM will enable us to adapt these, so they are consistent. We will do so to the minimum degree possible while also feeding back to amend the source material, with evidence-based testing, on actual digital twins and their operational data. We will need to ensure our community is not incentivised (by academic credit or profit) to do “green fields” work rather than “adopt and adapt”: we will seek authoritative sources for our reference data. Much work has already taken place in developing interoperable data exchange formats and classification standards that can describe the built environment at scales from a component of a video camera to an entire facility such as an airport. These data exchange formats have already



been adopted by software vendors, and we will look to adopt these international standards, and seek to influence their ongoing development by engaging with their governing bodies. Knowing that the intention is to ultimately establish these formats as national standards, care needs to be taken to ensure that they reference and sit within existing standards that support these concepts.

Secure by design

Role-based access control, ownership and authorship information, and usage restrictions will be built in throughout the work. Information providers must have confidence that they control who has access to their data, with the option to dynamically control availability to different data, and adoption of appropriate security levels. The deliverables should be able to address security-related requirements, those related to the managed control of information that should not be released openly or has handling caveats, and those requirements around the integrity of the implementation so that systems that make use of them can be trusted to handle the information.

Examples of areas that have such requirements include data that has personal, commercial and national security sensitivities as well as fundamental needs to trust whatever data is created and managed by DBB itself - a mix that is common in infrastructure-related applications. This approach is essential to ensure trusted

adoption in industry and government and should be seen as complementary to the “default to open” message. All deliveries and applications should benefit from an approach that intends to be trustworthy. Security by design applies not only to individual components of the architecture, but to the integration architecture as a whole. Adopting a systems-of-systems approach is essential to address security issues (such as access control and data aggregation issues), across the NDT ecosystem.

Open for participation

We will emphasise the use of openly available tools and standards in defining the Commons. By proposing and standardising open protocols we will minimise barriers to participation in the Commons arising from vendor lock-in. We must be able to freely distribute new content developed in the Commons, and new actors will be able to contribute data sets and digital twins to the NDT with zero cost of entry.

Many uses of the eventual NDT will require users to invoke a computationally expensive model component, or to access high-quality data with a cost. However, we will curate publicly documented exemplar models, both to provide examples of how to use the framework for contributors, and to prove the Commons work is fit for purpose. While we will endeavour to use existing open source tools from the community, rather than creating tools ourselves, if improvements or amendments to tools or standards are

required to make them appropriate to our project, such changes will be contributed back “upstream” to the community through appropriate “pull requests”. Any tools and data created in the Commons will be released using business-friendly open source licences, with mechanisms to balance the cost and benefits of upgrading/updating data.

However, we should also not use that subset of open source components with “viral” licences that would be hostile to attracting businesses to the NDT ecosystem: our outputs must be able to be incorporated into open or commercial software. We prefer the permissive-licence sub-sector of the open community.

Focus on quality

We know where we want to go: a world where data and models are consistently described, updated and applied across multiple sectors. We should keep our eyes on the prize and not give up on this goal in the face of messy reality.

In developing the FDM and RDL we will use a top-down and bottom-up approach respectively, ensuring that they join in the middle. The FDM will draw on established concepts from philosophical ontology and industry-level data models, making choices between alternative treatments of real-world phenomena where necessary. The aim is to build an ontology that allows us to state everything that can sensibly be stated, but each in only one way. Our foundation

data model should, as early but as richly as possible, describe “life, the universe and everything” [8], while the reference data library will grow over time as our scope increases.

The RDL, by contrast, will be drawn initially from existing authoritative sources, both harmonising across sources and integrated into the FDM both confirming (or requiring extension to) the coverage of the FDM, and confirming the quality or identifying the need for its improvement in the RDL. Further checks will be made against our example digital twins: that their content can be mapped into the FDM and RDL. Our foundation data model should, as early but as richly as possible, describe “life, the universe and everything” [8], while the reference data library will grow over time as our scope increases, both developed alongside real-world digital twins.

It is in the nature of a FDM that it takes a large part of it to say anything significant, which is the flip side of it enabling anything valid to be said. Our aim is to deliver a minimum viable product that we can be confident it is extensible while minimising the risk of the need to make changes to the work already done.

We know it will not be possible to have all builders of digital twins use the FDM and RDL for their creation. So, the FDM and RDL must be of sufficient descriptive power that others can transform their digital twins into ones that are compliant.


4.2. Work Programme

The following tasks describe a multi-year programme of work to deliver an initial version of the Commons.

These tasks will deliver a set of proposed protocols for standardisation. In the process, they will also create a working pilot system that proves the validity of the protocols proposed, together with the technical collateral (documentation, review processes, testing and validation tools) needed to underpin the processes for governance of the NDT.

Finally, they will start to bring together the teams, cultures and processes that will sustain the NDT in the longer term. We know that alone a series of task-and-finish activities will not produce a sustainable long-term NDT, but we hope that the work will produce the beginnings of an organisation that could do so.

The focus of these tasks will be on creating an evidence-based design that meets the requirements set out here. The preparation of published reports and papers is proposed as a mechanism to encourage the delivery of the technical capability: we note that this documentary evidence of progress is not an end in itself.



We know that alone a series of task-and-finish activities will not produce a sustainable long-term NDT, but we hope that the work will produce the beginnings of an organisation that could do so.

Initial tasks to establish the Commons

Our task is to build a system that can answer questions that require multiple twins to be integrated: by using the results or data from multiple twins to answer a question or as input to some analytic process.

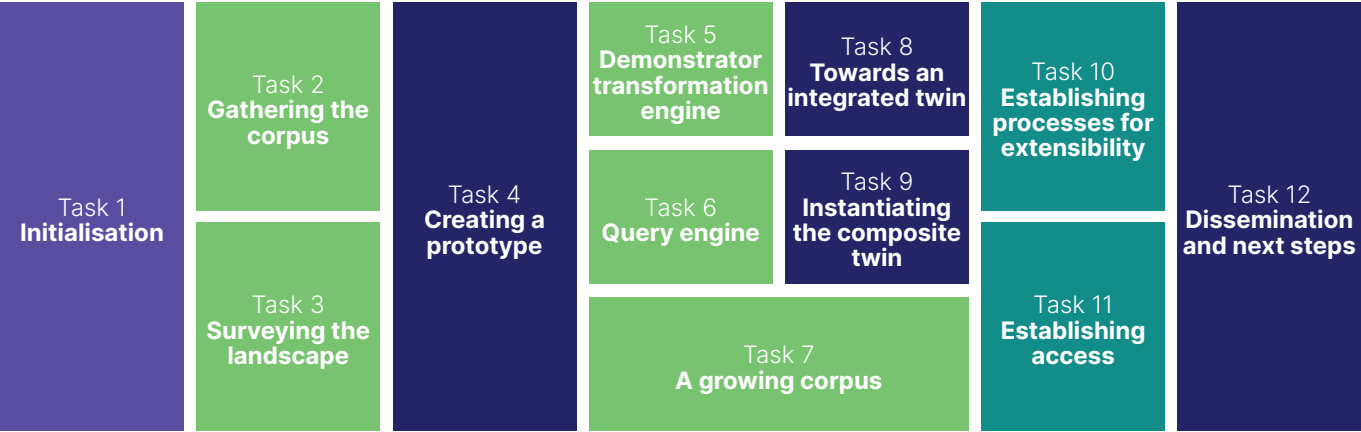


Figure 4: Tasks to be completed in the development of the Information Management Framework for the National Digital Twin

Task 1: Initialisation

This task prepares the ground for the technical work to follow: appointing leaders and teams, building governance structures and establishing ways of working (including the version control and project management tooling for the programme of work). We will establish the processes and criteria for review and acceptance of work package deliverables, including the mechanisms to be used to address any non-conformance. We will work with the Enablers and Change streams to establish stakeholder engagement process and independent testing teams and scenarios. We will be working with the Governance stream to establish ongoing oversight of the programme to monitor both progress and “direction of travel”, to address emerging issues and challenges and to revalidate the overall programme and approach at the end of each task and at key milestones, including standards development.

As security needs to be baked into the integration architecture developed in subsequent work packages, the initiation work package should ensure that oversight and planning will put in place independent scrutiny of security requirements. This should include links from other workstreams.

Task 2: Gathering the corpus

In this task, we will focus on gathering and understanding a collection of digital twins, which will form part of the evidence basis to check our data model and RDL; derived from the founding members of the DT Hub, the work of DAFNI and other sources. In addition to even coverage of the sectors that will be addressed in our composite digital twin, we will also look at some digital twins outside our scope, so we can confirm our foundation data model is robust to increases in scope.

Differences in the source models will reflect a variety of different ways things can be viewed, which would, if used unaltered, impose varying constraints on what data can be held, in particular outside the scope of the data in the twins being analysed. Our parallel work in Task 4 will need to define a principled data model that is capable of supporting the data requirements of all of these. Methodologies for re-engineering ontologies from source data should be considered (for example, the work of [9]).

Deliverables: a set of digital twins for use in developing further deliverables, and a published pre-print, highlighting commonalities and differences between approaches and data structures in the twins in this corpus.

Task 2B: Assessment of security threats and approaches

In this task, we will review security approaches, (considering PAS 185 and ISO 19650-5) and threat models. Although security-related information will be embodied throughout the resulting work, the security significance of the NDT means this will need to be considered upfront, and the results shared with those carrying out the remainder of the tasks, to ensure their awareness of the security context. Threat models will vary considerably depending on the types of application and digital twin involved.

In considering the security issues, it is necessary to look beyond the traditional concepts of confidentiality, integrity and availability to assess and understand the implications in respect of:

- **authenticity** – is the data/information being used of known and acceptable provenance and how can this be verified?
- **utility** – the asset will change over its life cycle, through maintenance or replacement or components or augmentation to change its capability or capacity. How will one establish that the historic telemetry data is comparable and consistent with that currently being used, and if not, how are long term trends monitored and managed?
- **safety** – what effect do changes to the physical and digital elements of the twin have on the safety case for the asset, its use or outputs?

- **resilience** – of the digital twin, how will it transform, renew and/or recover, in a timely manner in response to adverse events?
- **control** – will access to the digital twin enable unauthorised control or modification of the physical asset, or enable hostile reconnaissance?
- **sensitive information** – how will the infrastructure prevent transfer and release of personally identifiable and commercially sensitive information, and how will the intellectual property of twins be protected?

We will be required to handle situations where a query requires access to data that the user is not authorised to use. We must consider how to respond to an authorisation failure in circumstances where it would be inappropriate to inform the user of the authorisation failure, as this would acknowledge the presence of information that they are not authorised to access: proposals on how to respond to this challenge should be made here. (For example, one common choice is to return the “wrong” error code, returning “not found” where the “correct” response is “forbidden”.)

Deliverable: A limited distribution paper discussing the threat modelling and security architecture of the proposed approach.



Task 3: Surveying the landscape

In this task, undertaken in parallel with Task 2, we review the existing formal and informal ontological standards in related spaces (nationally and internationally): infrastructure, utilities, building information modelling, NBS, smart cities, semantic specifications of mathematical models, and security and access control models. Particular emphasis should be placed on finding authoritative and well-accepted sources for this: if standards have been consistently adopted by communities, we should build on these. We will review choices adopted by major infrastructure organisations knowing that these may, in themselves, be de facto standards. In assessing the consistency or otherwise of available existing work, the adherence both to format and structure will be assessed alongside the actual characteristics and quality of the data available.

This task will also survey the landscape of available tools for editing and checking ontologies.

This work should include a survey of available upper ontologies, such as [9], [10], [11], and that in [6]. Recommended standards for expression and documentation of upper ontologies should be reviewed, including [12].

This task will also review formalisms and structure that can be used for organising ontologies within the programme for particular purposes, such as [13] and [14].

Deliverable: A collection of industry ontologies and top-Level ontologies. Standards for documenting ontologies, and languages for representing them. A series of survey papers, describing the commonalities, differences and approaches taken in the ontologies and tools discovered. This should also identify what is missing. The report should set out options where multiple existing pieces of work are not compatible [15], [16].

Task 3b: Initial determination of core choices

Based on the work of Task 3, the NDT programme will make an initial selection for some of the options for design of the information management framework, particularly where such a decision needs to be made before work on Task 4 can begin. We will work closely with the enablers and change streams to highlight, check and sell the decisions being made. Where there are competing viewpoints, we will explore the dimensions of disagreement to find a robust way forward. These decisions will be revisited in light of the outcome of later work packages, but once made, there will be significant sunk costs, so a careful decision at this stage is necessary.

In particular, we must base our ontological work on a firm and consistent footing by choosing an appropriate formal “top-level ontology” to begin to work with. This must have the broadest possible scope, to establish our ability to modify and extend our programme while retaining consistency.

Task 4: Creating a prototype

In this task, work will begin on adopting, adapting and developing the foundation data model and collecting and validating the reference data library. To focus this work and place it in context, the data model and reference data will be tested and iterated by considering four digital twins from our corpus. While maintaining an unchanged copy of the primary source, we will build a two-way mapping into our common semantic description.

To do this, first, the data structures and their intended meanings will be considered, and how they can be supported by the FDM and RDL, for example by including appropriate classes in the RDL. Second, the data will be analysed to identify uses of the data structure that do not conform to the intended usage. This will be a significant task, especially when the twins are industrial scale.

Our FDM must be able to describe the upper parts of our specific problem domain: proposing the generic objects and relationships to be held within and between digital twins, models, data sets, and physical twins. This work will draw upon careful consideration of industry data models used in the corpus developed in Task 2.

It is likely that we will discover things about the source twins that will hamper this, and it will be necessary to address these as part of the process, or to develop ways in which

defects can be overcome. We will use the full corpus gathered in Task 2 to check that the data model and reference data library meet our data requirements internally, and, when we are ready, engage with external stakeholders across the DT Hub and elsewhere to validate this work, in Task 4b.

We will use domain ontologies discovered in Task 3 as input to this work, quality assuring these as necessary to be consistent with the common upper ontology. This formal description should be able to be automatically transformable back into the individual data files and software systems of the existing digital twins (except to the extent that the original data may be inconsistent such that no such reconstruction can occur.) The reader should note that we do NOT propose to use this to back-propagate to the twin owners. (Quite reasonably, we do not expect this would be acceptable to stakeholders!) However, the in-principle ability to transform back and forth between data models and reference data, is an important validation of the technology.


In this task, we will also deploy databases, objects, or triple stores as needed to host these models. In doing so, we will create the origins of our discovery protocols. This will also include the data needed for a role-based access model: models of organisations, threats, actors and their roles. This will be more than just a controlled vocabulary: the nature of the relationships between such entities will be modelled

based on their interdependency. The authentication systems and application security for our proof-of-concept databases will be constructed according to this model.

Developing this proposal will require selecting and using tools for editing and checking ontologies, such as [17].

Task 4b: External review of the model

In parallel with Task 4, and working through the Enablers stream and DT Hub, we will engage with a small sub-set of stakeholders, including asset owners and internationally renowned experts, to further test the work, beyond the reference prototype composite twin being used there. This group will both provide a useful forum for the Task 4 team to test and resolve challenging areas and pass challenges identified by this work package back to Task 4 for resolution. They will attempt to independently validate the work in development by trying it with their own use-cases: for example, this could include a group working on using the framework with digital twins related to major transport investments in their impact. Other examples could consider utility provision to a major industrial park, transport/pedestrian modelling relating to a retail area/university campus. This work will need to surface whether these stakeholders find the security and discovery models fit-for-purpose, as well as the models for describing the twins themselves.



Our task is to build a system that can answer questions that require multiple twins to be integrated: by using the results or data from multiple twins to answer a question or as input to some analytic process.

It is critical that this group is made up of friendly critics: not afraid to give us difficult feedback, but supportive of the aims of the project as a whole. In this way, we will be better able to take other people with us as we progress the work.

The costs estimated for this work package include only the costs of coordination, dissemination and response: we hope asset owners will volunteer their own resources for their review.

At the end of this phase, with validation both against the gathered corpus and through the review group of this work package, we will publish our FDM and our prototype mappings as a technical White Paper, requesting comments. Beyond this stage, all proposals for tasks are highly speculative, contingent on the outcomes of Tasks 1–4 and support from the wider community.

Task 5: Creating a demonstrator transformation engine: automating data ingress

Based on the manual wrangling process of Task 4, we will survey and commission for the Commons semi-automated tools capable of replicating that transformation process and improving real-time ingress, perhaps making feature contributions to existing tools or building new ones if necessary. For example, we will commission intelligent agents that transform and input third-party spreadsheets into our semantic description, with the minimum of supervision. Extract-transform-load tools for structured data are established in software development practices, and some may be relevant to this task, but while our focus can be on structured or semi-structured information, we also need to assimilate less or unstructured data. Automatic knowledge extraction from unstructured sources is a key enabling transformative technology from AI.

A challenge that may arise is that some source models may have used different reference data libraries than we would recommend (or incorrectly used the RDLs we recommend, or proprietary ones we cannot access, or no standard RDL at all), especially when classes overlap rather than match. We will develop tools to semi-automate the process of transforming the source data to match our chosen RDLs. This will be in addition to performing a structural mapping.

(In the medium term, we would also hope to be able to influence upstream sources to adopt our data library choices to alleviate this issue.)

This work will allow us to establish the protocols and standards that will be needed to describe such services for transformation and delivery of data. The role of the Commons is to facilitate the development of a rich ecosystem of such services, using consistent protocols allowing for sharing and integration. We emphasise that the particular transformations we will construct are there to provide the scaffold to help us learn and deliver these protocols, not outputs in themselves.

As part of our security considerations, at this stage we will need to consider the threat of poor-quality data, including malicious data being proposed for integration.

In this task, we will also begin to develop tools to automatically validate whether the models are being used correctly. While it is possible to use the FDM only at the level of data structures, to validate correct use of the FDM, we will construct the rules that the data needs to comply with – an axiomatisation of the FDM. This supports the mapping of data into the FDM by preventing invalid mappings.

Task 6: Query engine

In the previous task, we considered how to ease the process of transforming digital twins which do not use our data models into the form of those models. In this task, we consider the protocols which will allow questions to be asked of digital twins in a consistent fashion.

Using the FDM and RDL describing digital twins for our reference examples, we will construct a system capable of answering questions by “running” (invoking or executing) a digital twin.

In some cases, the twin itself will be purely descriptive: in which case the necessary query engine will be akin to a (potentially distributed) database query.

In other cases, the twin will reside within its own computational infrastructure. The necessary query interface will therefore require an understanding of the necessary information envelopes for executing a twin, and obtaining it from information supplied in or available to the query engine. Software can check that the information available for an object is sufficient.

Another case is where the information about the twin, expressed using our FDM and RDL, will be sufficient that the twin could be invoked, but the necessary computational infrastructure is not already provided. For example, a twin could be described by giving the structural data necessary to run

a finite element model (FEM) of a bridge, but without the simulation tool capable of solving FEMs. In this case, we will need to implement the code to transform the data from our canonical form into the input format expected by the simulation code, and to formulate the message payload that carries that information to an appropriate simulation service. (The available services having been described in canonical form in the data store.) The security implications of which data can be accessed by whom, and sent to which analysis services, will need to be carefully considered, and the approach developed in Tasks 2 and 3 tested.

We note that this work does not require us to build the simulation service itself. Several already exist, for example in the form of various “science gateways” [18]. Actually, running dynamic models will not form part of our framework: rather, our query engine will know how to extract the necessary data from the backbone and pass it to execution services.

Our output will be the protocols and data standards allowing the flourishing of an interoperable ecosystem of such services. As a by-product, necessary to proving this system, we will construct a demonstrator of an analytical engine founded on the protocols and standards and able to enrol and inspire both users and suppliers of data and of digital twins.

Task 7: A growing corpus

As a background activity, additional digital twins will be being gradually added to the corpus we are aware of. As we become aware of them, we, both the core team and the team of stakeholders engaged through Task 4b, will continue to validate and challenge our work, by verifying that it is possible to map them onto the FDM and RDL, and identify any extensions required.

Task 8: Towards an integrated twin

Thus far, we have considered individual twins in isolation, using a common data model. With this work complete, we can now start to construct a composite digital twin. We will select two to four twins from the corpus, with a particular eye to a demonstrator of a linked super-twin, combining component twins at different scales, and capable of answering useful questions using the linked capabilities of more than one model. Simple linking of data might be somewhat trivial (e.g. taking a location from a building model and transferring that to Google Maps). Our task is to build a system that can answer questions that require multiple twins to be integrated: by using the results or data from multiple twins to answer a question or as input to some analytic process.

Using the FDM and RDL ontology, and the automated tools selected or developed in Task 5, we will, at this stage, describe the super twin, and the necessitated data interchange between components. In doing so, we will discover required extensions to our reference data library.

We may also discover problems with our FDM. Any issues found will be carefully logged and tracked: many will turn out to be “documentation bugs”: sound ontological choices that are harder to explain and use correctly. Others might require amendments, as opposed to extensions,

risking backward compatibility. Any amendments made at this stage should be against carefully labelled versions of the ontology. This may require the revision of existing mappings into the FDM/RDL and the re-running of the transformation of source data into compliant form with automated processes where possible.

At the end of Task 8, we will have developed the confidence in the correctness of our FDM and RDL to begin the release process of the first part of our reference data library. We will begin the process of formally standardising these as technical specifications, locking down the FDM against non-backwards compatible changes and encouraging their use within the user community in industry and government.

Task 9: Instantiating the composite twin

We will establish, as a reference example of what a complete digital twin ecosystem will look like, the services necessary for querying across, and iteratively communicating between, the components of the multi-scale twin.

This distributed reference implementation of the protocols and data sets necessary for a real, composite, multi-scale digital twin, will be a milestone for the project.

Task 10: Establishing processes for extensibility

As the Commons develops, we will receive proposals for contributions to extend the FDM, RDL, and for additional services for query execution or data transformation. We will need to establish processes for checking, approval and acceptance of these. This work will be carried out in close liaison with the Governance stream of the NDT project, with the Commons work focusing on the technical analysis of validity and consistency.

We will also receive proposals for external data sources that would like to become discoverable through the protocols defined in the Commons, with appropriate (real-time or intermittent) mappings to our FDM and RDL. Again, we will need to establish processes to ensure that the source quality, and the quality and reliability of the mapping, are sufficient. Again, we will establish the appropriate technical quality standards and the review of security implications to facilitate this review, and liaise with the Governance stream to formalise and socialise the processes by which these are used.

Alongside developing the RDL, we will need to develop/document the methodologies and standards for extending them, since these will have become apparent in executing this task.

Task 11: Establishing access

Throughout the preceding tasks, we will, by necessity, and in conjunction with the stakeholder testing community established in Task 4, have evolved scripting interfaces and domain-specific languages enabling the data-driven expression of queries to the framework. In this task, we will polish these for use outside the project community, creating a user-friendly web-based query interface. As we prepare to bring in users outside the extended team of close early adopters, we will need to firmly validate that our authorisation engine is robust to engagement from outside the consortium: additional penetration testing will be needed as part of this work.

Task 12: Dissemination and next steps

We will document our work and conclude the process to turn our technical specification into a national standard. In particular, we will develop and share materials describing what is required by third parties to publish digital twins and reference data consistent with our framework. We will work closely with the Governance team to identify how verification of compliance will be undertaken and delivered, and we will develop material that defines what compliance means.

References

- 1 Centre for Digital Built Britain (2018) Gemini Principles. <https://doi.org/10.17863/CAM.32260>
- 2 National Infrastructure Commission (2017) Data for the Public Good. <https://www.nic.org.uk/wp-content/uploads/Data-for-the-Public-Good-NIC-Report.pdf>
- 3 Hackitt (2018) Building a Safer Future: Independent Review of Building Regulations and Fire Safety. Final Report. <http://www.gov.uk/government/publications/independent-review-of-building-regulations-and-fire-safety-final-report>
- 4 Mosavi, Ozturk and Chau (2018) Flood prediction using machine learning models. Water. 10(11), p. 1536.
- 5 Hey, Tansley and Tolle (2009) The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA: Microsoft Research.
- 6 West (2011) Developing High Quality Data Models. San Francisco, CA: Morgan Kaufmann Publishers Inc. West (2011) Developing High Quality Data Models. San Francisco, CA: Morgan Kaufmann Publishers Inc.
- 7 Mordue (2015) BIM Levels of Information. NBS. Available at: <https://www.thenbs.com/knowledge/bim-levels-of-information>
- 8 Adams (1980) The Hitchhiker's Guide to the Galaxy. New York: Harmony Books.
- 9 de Cesare and Partridge (2016) BORO as a Foundation to Enterprise Ontology. Journal of Information Systems. 30(2), pp. 83-112. <https://doi.org/10.2308/isys-51428>
- 10 International Organization for Standardization (2003) Industrial Automation Systems and Integration - Integration of Life-Cycle Data for Process Plants Including Oil and Gas Production Facilities - Part 2: Data Model (ISO 15926-2:2003).
- 11 Arp, Smith and Spear (2015) Building Ontologies with Basic Formal Ontology. The MIT Press.
- 12 International Organization for Standardization (2019) Information Technology - Top-Level Ontologies - Part 1: Requirements. (ISO/IEC DIS 21838-1:2019).
- 13 Antoniou et al. (2003) Web Ontology Language (OWL). In Handbook on Ontologies, edited by Staab and Studer, pp. 67-92. Springer.
- 14 International Organization for Standardization (2007) Information technology. Common Logic (CL). A framework for a family of logic-based languages. (ISO/IEC 24707:2007).
- 15 International Organization for Standardization (2020) Industry Foundation Classes (IFC) for data sharing in the construction and facility management industries. Data schema. (BS EN ISO 16739-1:2020)
- 16 Open Geospatial Consortium and buildingSMART International (2020) Built environment data standards and their integration: an analysis of IFC, CityGML and LandInfra.
- 17 Musen (2015) The Protégé Project: A Look Back and a Look Forward. AI Matters, 1(4), pp. 4-12. doi:10.1145/2757001.2757003.
- 18 Kacsuk (2014) Science Gateways for Distributed Computing Infrastructures: Development Framework and Exploitation by Scientific User Communities. Springer.

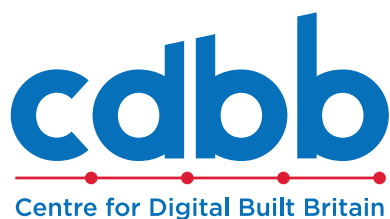
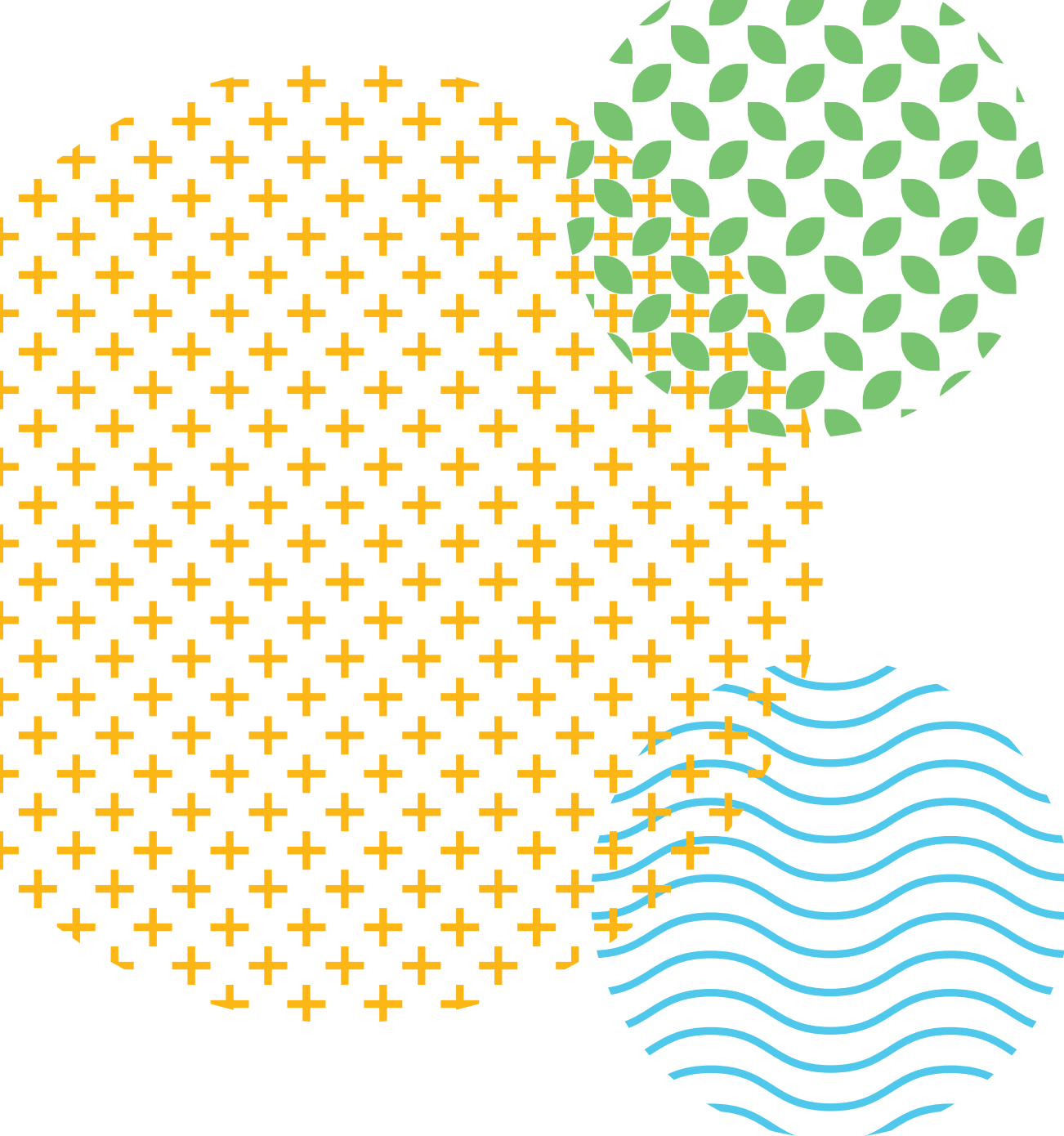
Acknowledgements

Named Authors:

James Hetherington, The Alan Turing Institute
Matthew West, Information Junction

Contributors:

Segun Alaynde, Heathrow	William Lopez Campo, EY
Karen Alford, Environment Agency	Alexandra Luck, A Luck Associates
Barry Blackwell, BEIS	Chara Makri, CDBB
Alexandra Bolton, CDBB	Richard Mortier, CDBB
Patrick Bossert, EY	Casey Mullen, Bentley Systems
Charles Boulton, Charles Boulton Ltd	Alex Murray, IPA
Hugh Boyes, Bodvoc Ltd	Alan Muse, RICS
Daniel Braund, Sellafield	Nick Nisbet, AEC3
Johanna Brown, Google	Ron Oren, Connected Places Catapult
Amelia Burnett, CDBB	Peter Parslow, Ordnance Survey
Sarah Campbell, UKRI NERC	Chris Partridge, BORO Solutions
Mark Casey, UKHO	Hugh Phillips, The Geospatial Commission
Lawrence Chapman, TempleGate Projects Ltd	Anna Radford-Watt, Mott MacDonald
Phani Chinchapatnam, Network Rail	Yacine Rezgui, Cardiff University
Samuel Chorlton, CDBB	Clive Roberts, University of Birmingham
Anthony Cohn, University of Leeds	Andrew Robinson, Urban Innovation Labs
Al Cook, Critical Insight	John Robison, Sellafield
Joel Crook, Google Cloud	David Rogers, HS2
Jon Crowcroft, The Alan Turing Institute	Dan Rossiter, BSI
Eric Daub, The Alan Turing Institute	Michael Rustell, Brunel
Sarah Delany, NBS	Greg Schleusner, HOK
Peter Demian, Loughborough University	Jennifer Schooling OBE, CSIC
Jozef Doboš, 3D Repo	Julian Schwarzenbach, IAM
Tim Drye, Yellow Zebra Artificial Intelligence	Miranda Sharp, Ordnance Survey
Peter El Hajj, Mott MacDonald	Dennis Sheldon, Georgia Tech
Ray Ellison, CIOB	Olivier Thereaux, Open Data Institute (ODI)
Mark Enzer, Mott MacDonald	Neil Thompson, Atkins
Simon Evans, Arup	Zane Ulhaq, Atkins
Paul Fjelrad, OFGEM	Liz Varga, DAFNI
Saadia Hakim, Google Cloud	Benjamin Walden, The Alan Turing Institute
Colin Henderson, Atkins	Jeremy Watson CBE, BRE
Paul Hodgson, Greater London Authority	Steven Yeomans, BRE
Rollo Home, Ordnance Survey	
Nadir Ince, GE Power	
Phil Jackson, IntraTeamIT Consultants Ltd	
Caroline Jay, University of Manchester	
Kell Jones, UCL	
Branwen Kelly, SNC-Lavalin	
Anne Kemp OBE, Atkins	
David Leal, Caesar System	
Thomas Liebich, AEC3	



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 & Innovation. This paper was also supported by the Construction Innovation Hub with funding provided through the Government's modern industrial strategy by Innovate UK, part of UK Research & Innovation.

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

