Future Generation Computer Systems 152 (2024) 112–126

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

A derived information framework for a dynamic knowledge graph and its application to smart cities

Jiaru Bai [a], Kok Foong Lee [b,c], Markus Hofmeister [a,c], Sebastian Mosbach [a,c], Jethro Akroyd [a,c], Markus Kraft [a,c,d,e,*]

[a] Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge, CB3 0AS, United Kingdom
[b] CMCL Innovations, Sheraton House, Cambridge, CB3 0AX, United Kingdom
[c] CARES, Cambridge Centre for Advanced Research and Education in Singapore, 1 Create Way, CREATE Tower, #05-05, Singapore, 138602, Singapore
[d] School of Chemical and Biomedical Engineering, Nanyang Technological University, 62 Nanyang Drive, Singapore, 637459, Singapore
[e] The Alan Turing Institute, London, NW1 2DB, United Kingdom

ARTICLE INFO

Keywords:
Dynamic knowledge graph
Derived information
Data provenance
Directed acyclic graph
Smart cities

ABSTRACT

In this work, we develop a derived information framework to semantically annotate how a piece of information can be obtained from others in a dynamic knowledge graph. We encode this using the notion of a "derivation" and capture its metadata with a lightweight ontology. We provide an agent template designed to monitor derivations and to standardise agents performing this and related operations. We implement both synchronous and asynchronous communication modes for agents interacting with the knowledge graph. When occurring in conjunction, directed acyclic graphs of derivations can arise, with changing data propagating through the knowledge graph by means of agents' actions. While the framework itself is domain-agnostic, we apply it in the context of smart cities as part of the World Avatar project and demonstrate that it is capable of handling sequential events across different timescales. Starting from source information, the framework automatically populates derived data and ensures they remain up to date upon access for a potential flood impact assessment use case.

1. Introduction

Inspired by Semantic Web technology, knowledge graphs are gaining popularity both in enterprise applications [1] and research fields [2]. They are seen as a suitable approach to integrating diverse information sources and fostering common understanding among domain experts [3,4]. Some renowned examples of static knowledge graphs are DBpedia [5] and Wikidata [6].

The dynamic aspect of knowledge graphs has recently been studied [7,8], where they are used as hubs for integrating complex systems of software agents, connecting different domains in specific use cases. This allows for what-if scenario analysis and automated decision-making that mimics human behaviour. This is a step towards the original vision of the Semantic Web, which is a fully annotated web of machine-readable data that can be processed autonomously by software agents [9,10]. However, this also highlights the need for robust methods to manage the changes and interdependencies in the complex information network, especially in a data-rich world which is rife with misinformation.

A key enabling factor identified by the community is provenance [11], *i.e.* where a piece of information originates from and how it came about. Provenance covers many aspects, including published literature, the wider internet, or data directly acquired through a measurement device. A number of provenance ontologies have been developed in the literature, for example, PAV [12], W3C Provenance Ontology (PROV-O) [13], Dublin Core Terms (DC Terms) [14], Bibliographic Ontology (BIBO) [15], and Open Provenance Model Ontology (OPMO) [16]. For a comprehensive review of developments in provenance data models, interested readers are referred to Sikos and Philp [17].

In dynamic knowledge graphs, one can consider a specific sub-problem of provenance, namely when some pieces of information are directly calculated from, or derived from, other pieces of information by software agents, all of which are already stored in the knowledge graph. When multiple pieces of information depend on each other in a certain way, the resulting cascade of information is progressed in time via a series of coordinated agent communications, where the calculated values of one process are inputs for other subsequent processes. To

* Corresponding author at: Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge, CB3 0AS, United Kingdom.
E-mail address: mk306@cam.ac.uk (M. Kraft).

https://doi.org/10.1016/j.future.2023.10.008
Received 19 May 2023; Received in revised form 29 August 2023; Accepted 9 October 2023
Available online 29 October 2023

make the knowledge graph truly dynamic, the system should also enable automated propagation of perturbations in source input data. This introduces another point of view — caching. In addition to providing functions (agents) to calculate a result, that result is stored, *i.e.* cached, in the knowledge graph. Many of the technical challenges are similar, such as being able to detect when the cache is out of date, and providing a function to update, refresh, or recalculate the cache.

Any approach to solving this problem can benefit from the lessons learned in other fields. In scientific computing, such a composition of computing tasks is referred to as *workflow* or *pipeline* [18,19]. Workflows are typically expressed as directed acyclic graphs (DAGs), where the nodes represent tasks and edges represent data flows [20]. Each task can only be started if all its precedent tasks are completed. There are many different workflow management systems (WMSs) that can be used to orchestrate these tasks, ranging from traditional paradigms like Pegasus [21,22] and Kepler [23] to modern approaches like Apache Airflow [24], Nextflow [25], and Lightning AI [26], to name a few. Some studies have focused on adaptive workflows, where changes in the input data may be incorporated into computation on-the-fly to provide real-time responses to dynamic events [27]. However, these systems often rely on heterogeneous and unstructured data models without semantic annotations [28], which can make it difficult to achieve interoperability across different systems, impeding the unification of the workflow community [29].

Another area which we may learn from is microservice architecture [30]. It is a service-oriented architecture (SOA) that emphasises loose coupling and high cohesion. It involves having each service in the ecosystem be independently developed, deployed, and maintained by a small dedicated team. When an *event* occurs, which is a similar concept as an execution instance of one pre-defined workflow, microservices are composed and coordinated via message-passing. This architecture places no restrictions on developers regarding the technology used to implement each microservice as long as a unified interface is agreed upon. Nonetheless, it places significant demands on the network to ensure successful communication. In this paradigm, Distributed Application Runtime (Dapr) [31] uses a sidecar to simplify communication via direct or event-based publish/subscribe messaging, unifying both modes of communication on the same platform.

Taking notes from these advents, we may summarise and suggest a knowledge-graph-native solution. By design, data in the knowledge graph can be uniquely identified via Internationalised Resource Identifiers (IRIs). All the active agents share the same world-view once granted access privileges. Analogous to the *message bus* in the microservice architecture, communication between agents operating on the knowledge graph can be delegated via serving the correct IRIs. This eliminates the necessity for large peer-to-peer data transfer. It can be further combined with the idea learned in the scientific workflow to model computation dependencies as DAGs in the knowledge graph. By encoding agents' messages as provenance records, we remove the need for direct agent-to-agent communication and allow information to travel through the knowledge graph. This decouples the system and allows for a distributed ecosystem of agents without the need for them to be aware of each other's existence.

As a "digital twin" of the world that is realised as a dynamic knowledge graph [7], the World Avatar is thought to be an appropriate candidate for evaluating the implementation of this technology. As it stands, there is an interwoven network of concepts spanning temporal and spatial scales, extending from molecule [32] and chemical mechanism [33] to laboratory [8], cities [34], and even the national level [7]. The World Avatar is constantly evolving as it is maintained by a network of active agents who regularly input new data and update existing data. To effectively manage the actions of these agents and ensure accurate tracking of their activities, an overarching architecture is required. Given the wide spectrum of data and use cases encompassed by the World Avatar, a generic and robust implementation of such an architecture is imperative.

The purpose of this paper is to provide a proof-of-concept for a technology-agnostic implementation of a derived information framework for dynamic knowledge graphs. This derivation framework is a realisation of such a knowledge-graph-native architecture. It uses a lightweight ontology to mark up provenance, an agent template to standardise agent operations, and an automated framework to propagate information changes in a dynamic knowledge graph. The design aims to lower the entry barrier for researchers to model any real-world cascading events with minimum effort by providing a user-friendly template. The significance of this framework lies in its dual capability to not only track and document the calculation process but also to automatically re-execute the computation when accessing outdated information. In an era characterised by both complexity and an abundance of data, the framework enables automated integration of the rapid influx of new information, ensuring constant access to up-to-date insights into the subject of interest.

In particular, we demonstrate this through a flood impact assessment use case within the World Avatar project. The selection of the smart cities domain is motivated by its necessity for handling both fast and slow urban dynamics [35], making it an ideal context for showcasing the framework's capability to accommodate real-world events with varying frequencies. Flooding events, as well as other natural disasters, require fast decision-making and a holistic perspective to respond. Hence, having up-to-date information at all times is valuable. However, it is essential to underline that the framework's versatility extends beyond the smart cities domain. Efforts are currently underway to apply this framework to various use cases in other domains.

The presentation of this paper is structured as follows. Section 2 situates the work by discussing the related work and summarising the lessons we learned through the development of a dynamic-knowledge-graph-based digital twin of the world; Section 3 provides the technical details on the complete architecture; Section 4 exemplifies the versatility of the framework via a use case in the context of smart cities; and Section 6 concludes the work.

2. Related works

This section provides a brief overview of the relevant ontologies on provenance for workflows that served as sources of inspiration. Additionally, our experience in developing the World Avatar project contributes to the rationale behind the derivation framework proposed in this study.

2.1. Provenance for workflows

As the *de facto* standard, PROV-O [13] adopts three base classes for provenance descriptions, *i.e.*, `prov:Entity` for the things subject to description, `prov:Activity` for the events occurring over a duration that lead to transformation of the entities, and `prov:Agent` for the things responsible for carrying out activities. PROV-O also provides qualified terms as elaborated information on binary relations between these base classes. For example, `prov:Association` qualifies the relationship `prov:wasAssociatedWith` between `prov:Activity` and `prov:Agent` by pointing to an instance of `prov:Plan` using `prov:hadPlan`. This qualification indicates the steps of action undertaken by the agent to achieve its goals. However, `prov:Plan` is provided as a rather isolated concept, lacking further specifications on how it can be connected to other concepts that are relevant for execution. Hence, PROV-O appears more towards retrospective recordings of events rather than actively invoking agents to perform tasks.

P-Plan [36] ontology partially bridges this gap by extending PROV-O with terms `p-plan:Step` and `p-plan:Variable` in scientific processes. This expansion helps to publish the methods and processes of scientific workflows as linked data. This work is later combined with Open Provenance Model (OPM) [37], leading to OPMW [38] which supports the connection between a workflow template and its concrete

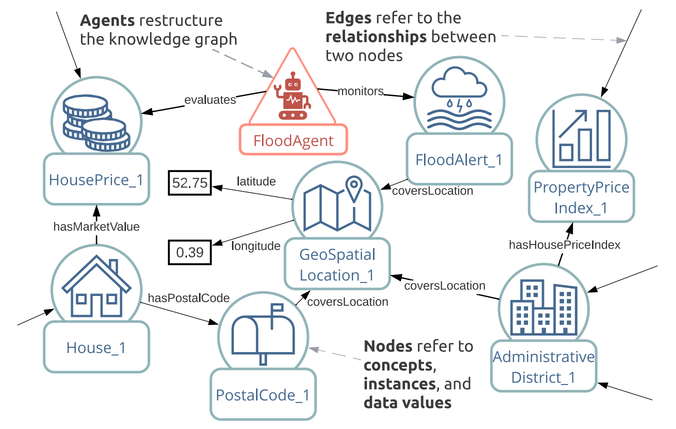


Fig. 1. Schematic of the World Avatar knowledge graph. Note that the figure is only illustrative and does not reflect actual data.

executions. Notably, OPM was a legacy data model developed from the Provenance Challenge series [39] and was actually the reference for the creation of PROV-O [38]. The authors of OPMW acknowledged that one aspect that is not yet supported by OPMW is the automatic re-execution of the published workflows in heterogeneous computing environments [38]. In that regard, modern containerised solutions and cloud services can be a potential solution [40]. The ideas behind these works served as inspiration for our knowledge-graph-native approach.

2.2. *The World Avatar*

In the World Avatar, the physical world we live in is captured and represented as the *base world*. The hypothetical versions of the world in which certain variables or assumptions are different are represented as *parallel worlds*. These alternative universes are managed by software agents that can perform a variety of tasks. Importantly, the World Avatar views these agents to be part of the knowledge graph, as depicted in Fig. 1. Using this technology stack, the World Avatar is versatile in three aspects: (1) answering cross-domain questions about the base world [41], (2) controlling real-world entities [8,42], and (3) supporting what-if scenario analysis with parallel worlds [43,44].

As the World Avatar project continues to evolve, it strives to more accurately represent complex phenomena across spatio-temporal scales. This requires a tool for efficiently coordinating the actions of agents that update and restructure the knowledge graph. Currently, software agents are represented similarly to semantic web services [45], which are invoked through HTTP requests. For time-consuming computations, an asynchronous job watcher is available to delegate jobs to high-performance computing (HPC) facilities. This was applied to assess the impact of quantum calculations on the air pollution dispersion [46] and to automate the calibration of combustion mechanisms [47]. These approaches largely adhere to the static remote procedure call paradigm. To fully unlock the potential of the dynamic world model, a more flexible and adaptive architecture that can facilitate autonomous interaction between agents and the knowledge graph is preferred.

The derived information framework described in this work offers such a solution as a first step towards revolutionising the agents' operations in dynamic knowledge graphs.

3. Methodology

This section provides an overview of the technical aspects of the derived information framework. We begin by introducing the ontology created for annotating the provenance markup, then presenting the agent template that developers can use as a starting point when developing new agents. Lastly, we discuss a client library, which can be used to manage the derivation instances.

3.1. *Derivation ontology*

We refer to "derivation" as the record for a singular occurrence of the fact that some pieces of information are derived or calculated from some other pieces of information. The term "derivation subgraph" is used to describe the subgraph of all derivation-related markup when a collection of interdependent processes is represented. Since the information in a knowledge graph is captured in the format of Subject–Predicate–Object statements (triples), storing the markup to capture this fact can be considered as a way of attaching arbitrary metadata to triples [48]. For that, generic solutions have been developed or are currently in development, such as reification [49] and W3C's RDF-star [50], which at the time of writing is still at the draft stage. While some implementations exist, *e.g.* Blazegraph's Reification Done Right (RDR) [49,51] or more recently GraphDB [52], at present these are not sufficiently widely supported for implementing the derivation framework in a technology-agnostic way, without tying ourselves to a particular product. Therefore, we choose to state the required metadata explicitly as triples and introduce OntoDerivation as a lightweight ontology to serve this purpose. We first explain the terminological component (TBox), followed by an example instantiation of the assertional component (ABox). The connection between OntoDerivation and OntoAgent [45], an ontology used to define the capabilities of agents, is further demonstrated by how they may be used in conjunction to govern agent actions.

The decision not to directly reuse concepts from PROV-O was made to facilitate the implementation for the specific aspect of provenance (and their updates) that we are considering in this work — how a piece of information is calculated from other sources and when it occurs. Specifically, for the following reasons:

- The need to accommodate diverse temporal scales in the response time with regard to activities performed by the agents. This differentiation is essential for facilitating automated updates of derived information through software agents, specifically concerning synchronous and asynchronous processes as the latter requires the recording of job status. This relates to the TBox level and will be discussed in Section 3.1.1.
- The requirement for a unified representation of timestamps in denoting the timeliness of both pure inputs and derivation instances, which will simplify the implementation of the updating algorithm in determining whether a derivation is outdated. PROV-O annotates these timestamps using different data properties with the range `xsd:dateTime`, leading to additional conversion for handling derivations instantiated from different timezones. This relates to the instantiation of provenance records and we discuss it in Section 3.1.2.
- Instantiating every piece of information as `prov:Entity` makes it impossible to distinguish concepts in the agents' input/output (I/O) signatures, let alone differentiate the capabilities of distinct agents in a complex derivation subgraph. One possible workaround is declaring all concepts in the I/O as subclasses of `prov:Entity`, however, this introduces an additional level of abstraction to basically all concepts in the domain ontologies, which is unnecessary in our opinion. We discuss our solution in Section 3.1.1 and elaborate on agents' I/O in Section 3.1.3.

3.1.1. *OntoDerivation TBox*

Fig. 2 depicts two types of derivation, *i.e.* synchronous (`Derivation`) and asynchronous (`DerivationAsyn`) to accommodate situations that respond in different timescales when a request is received. The synchronous mode communicates via the endpoint exposed by the software agent. It is thus faster and hence intended for applications demanding real-time responses. The asynchronous mode communicates exclusively through the knowledge graph. It has the advantage of recording each stage of the operation in the knowledge

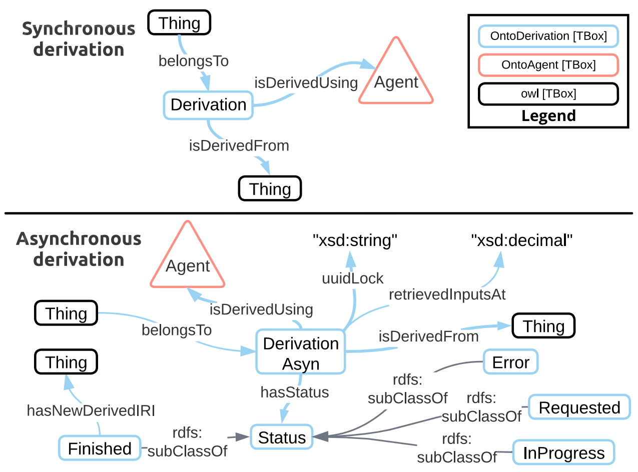


Fig. 2. Concepts and relationships of the OntoDerivation ontology. All classes and properties belong to the OntoDerivation namespace unless stated otherwise (for namespace definitions see Appendix A.1).

graph, but it is slower and hence better suited to relatively expensive jobs.

The information dependencies of a derivation are consistently marked regardless of the communication protocol, with the derived information (outputs) `belongsTo` the derivation, which itself `isDerivedFrom` some source information (inputs) and `isDerivedUsing` an agent defined in OntoAgent [45]. It is worth noting that there is no limit on the number of inputs or outputs for a derivation, but one output entity cannot `belongsTo` more than one derivation instance. Both source and derived information are abstracted using `owl:Thing`, so it is also possible for an input of a derivation to be part of another derivation instance, meaning that the input is a piece of derived information itself and `belongsTo` another derivation.

To support asynchronous operation, the concept `Status` is introduced to mark the state of an asynchronous derivation with the available options of `Requested`, `InProgress`, `Finished` and `Error`. The data property `retrievedInputsAt` records the timestamp when the inputs for the derivation were read in order to start the computation, which will later be used to update the timestamp for the derivation instance. The data property `uuidLock` uniquely identifies the agent thread that is processing the derivation and prevents any amendment from other threads that do not hold the correct key. This ensures thread-safe operations when multiple threads are employed to boost the throughput of derivation processing. A specific object property called `hasNewDerivedIRI` is used at the `Finished` status to temporarily link any newly derived information. These output entities will eventually be reconnected to the derivation instance after the agents clean up the status. Like the final outputs, there is no limit on the number of new entities that can be connected through `hasNewDerivedIRI` for each derivation instance.

A description logic representation of OntoDerivation ontology is provided in Appendix A.2. The OntoDerivation TBox is available at [53].

3.1.2. OntoDerivation ABox

Fig. 3 exemplifies an instantiated derivation instance and its simplified representation. Upon initialisation, each derivation is annotated with a timestamp following the W3C standard [54]. The Unix timestamp is chosen over `xsd:dateTime` to enable direct numerical comparison. In the rest of this paper, simple integers will be used instead of the actual (Unix) timestamp for readability. Over the lifecycle of a derivation, this timestamp is used to assess whether it is out-of-date. In this example, as the timestamp of derivation is smaller than that of

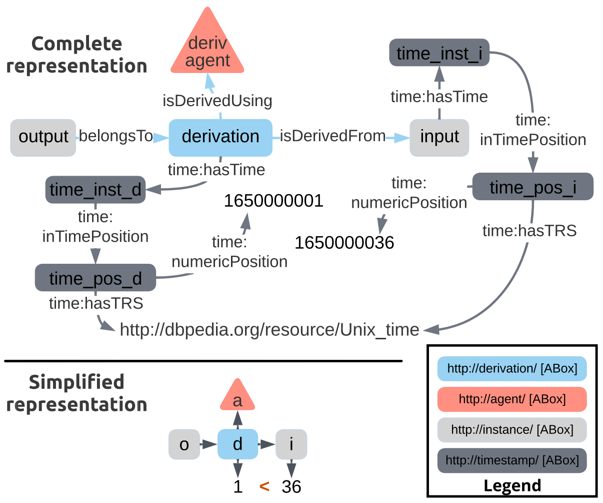


Fig. 3. An example derivation instance fully annotated with metadata and its simplified representation, which will be used throughout the rest of this paper. All properties belong to the OntoDerivation namespace unless stated otherwise (for namespace definitions see Appendix A.1).

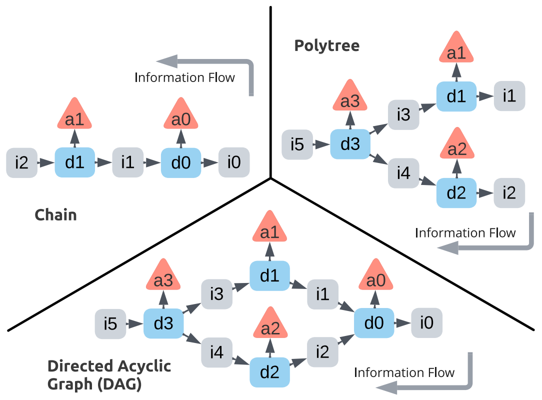


Fig. 4. Derivation subgraph structures as directed acyclic graphs (DAGs) of varying generality. The arrows between instances indicate the markup for data dependencies. The "information" flows in the opposite direction, *i.e.* from right to left.

its source information, *i.e.* $1 < 36$, we conclude that this derivation is outdated and that, hence, an update of the derived information is required. It should be noted, however, though the outputs of a derivation can be input to others, they are not directly associated with a timestamp (but are indirect via the derivations to which they belong). For example, the *output* instance in Fig. 3 is not directly associated with a timestamp and so are all other instances that `belongsTo` a derivation. This design choice is made to reduce the redundant information in the knowledge graph as the timeliness of the derived information is already reflected by the timestamp of the derivation instance it `belongsTo`. As a result, when a derivation subgraph is composed of several connected derivations, direct timestamp comparison becomes infeasible for those derivations whose inputs are outputs of another derivation instance. It is therefore necessary to employ additional criteria for determining whether a derivation is out-of-date. More information on its implementation is provided in Section 3.3.

Fig. 4 illustrates three levels of generality in a derivation subgraph, from basic linear *chain* to non-linear *polytree* to generic *directed acyclic graph*. Unlike in scientific workflows, where the arrows often point in

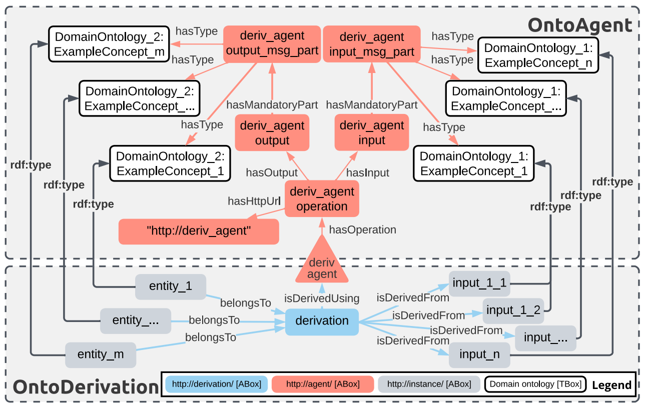


Fig. 5. The instantiation that connects OntoDerivation and OntoAgent. The linkage between derivation instances and agent instances is used to regulate agent operations. All object properties belong to the OntoDerivation namespace unless stated otherwise (for namespace definitions see Appendix A.1).

the same direction as the data flow, the markup in the knowledge graph denotes the data dependencies and points to the source of the information. The changes in the source information travel in the opposite direction within the knowledge graph, as illustrated by "information flow". In this context, we define 'upstream' and 'downstream' to refer to the relative location of a derivation instance within that flow. A key feature of the design is that only relevant downstream information is updated when accessed, which will be discussed in more detail in Section 3.3.

We emphasise that the primary intent of the derivation framework is to represent logical dependencies as such, as a historical record, rather than consider them as 'steps' in a workflow or algorithm. This implies that, in contrast to workflows [55], cyclic dependencies are not permitted in the derivation framework, as they would amount to logical contradictions. Nonetheless, the framework can of course be used to record the dependencies of pieces of information that were obtained from algorithms containing loops and other circular constructs.

OntoDerivation ontology is designed in a way that is easy to use and easy to query. Tools are provided to automatically generate derivation markup for all three types, and validate the generated derivation subgraph. An example SPARQL query to retrieve all derivations in a given knowledge graph is provided as Query 1 in Appendix A.3.

3.1.3. *Connection with OntoAgent*

In the World Avatar, OntoAgent [45] is used to mark up the I/O signature of agents to facilitate agent discovery. This markup points to concepts in domain ontologies and indicates the agent's capabilities by identifying the concepts required/produced by the agents. By contrast, OntoDerivation focuses on the instance level, *i.e.* actual data digested and produced by the agents corresponding to each occurrence of computation, revealing the opportunity for employing both ontologies to regulate agent operations.

Fig. 5 provides an example of instantiation using both OntoDerivation and OntoAgent, where inputs and derived outputs of the derivation instance are both instantiated from the I/O signature defined in the OntoAgent instance. Specifically, the relationship `OntoAgent:hasType` can indicate that an agent's message encompasses various concepts from domain ontologies. For a given derivation instance, the inputs can be classified into key–value pairs, with the IRI of each concept as the key and the list of instances as the value. For example, the value of the pair with the key *DomainOntology_1:ExampleConcept_1* will be *input_1_1* and *input_1_2*. This design connects the conceptual capability of the agents with the concrete tasks that they are assigned to execute. Following this practice, the development of agents can be focused on the concept level, simplifying the implementation.

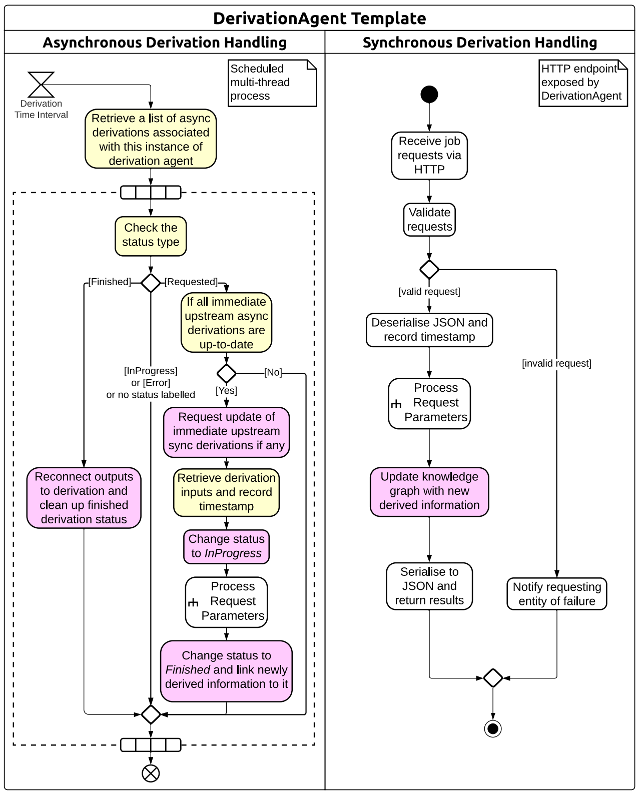

DerivationAgent Template
Asynchronous Derivation Handling
Synchronous Derivation Handling
Derivation Time Interval
Scheduled multi-thread process
Retrieve a list of async derivations associated with this instance of derivation agent
Check the status type
[Finished]
[Requested]
If all immediate upstream async derivations are up-to-date
[InProgress] or [Error] or no status labelled
[No]
[Yes]
Request update of immediate upstream sync derivations if any
Reconnect outputs to derivation and clean up finished derivation status
Retrieve derivation inputs and record timestamp
Change status to InProgress
Process Request Parameters
Change status to Finished and link newly derived information to it
HTTP endpoint exposed by DerivationAgent
Receive job requests via HTTP
Validate requests
[valid request]
[invalid request]
Deserialise JSON and record timestamp
Process Request Parameters
Update knowledge graph with new derived information
Serialise to JSON and return results
Notify requesting entity of failure

Fig. 6. UML activity diagram of the derivation agent template supporting both synchronous and asynchronous derivations. Developers need only supply the activity node `ProcessRequestParameters` for specific agent capabilities. The yellow- and magenta-shaded actions represent knowledge graph data retrieval and population operations respectively.

3.2. *Derivation agent*

The use of ontological markup to record each step in the process of updating derived information largely restricts the communication an agent needs to perform to the knowledge graph itself, rather than with other agents. We provide an agent template to support this in both synchronous and asynchronous modes. The template makes use of data container classes to host agents' inputs and outputs. These classes are key–value pairs that may be mapped using Query 2 in Appendix A.3. Developers are supplied with utility functions to access/validate the mapped inputs and to construct outputs. The agent logic that computes outputs from the inputs is the only code required from the developer. We made the template available in both Java and Python to increase accessibility and depict its unified modelling language (UML) activity diagram in Fig. 6. The essential design elements are elaborated on below.

3.2.1. *Synchronous communication mode*

The synchronous communication mode is realised through direct agent requests/responses. Upon receiving a request for a normal `Derivation`, the agent serialises the request content to an instance of the container class. The time instant is recorded immediately before passing the inputs to be processed by the developers' code. This instant is considered as the timestamp when inputs were read. Once the outputs are constructed, an update operation will be formulated and executed by the agent to update the knowledge graph. If there is no error, the derivation update is considered successful and a response will be returned that includes the produced derived information and the recorded timestamp.

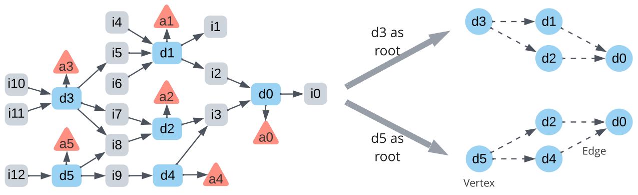


Fig. 7. DAGs are used in memory to assist the backtracking of the DFS algorithm when updating the derivation subgraph. The derivation instances are treated as vertices and their connections as edges. Depending on the root derivation chosen, different DAGs can be generated.

3.2.2. Asynchronous communication mode

In the asynchronous communication mode, agents monitor the status of derivations in the knowledge graph and perform any requested tasks. When an agent detects a derivation with the status `Requested`, it first checks if all the data dependencies for that derivation are satisfied using Query 3 in Appendix A.3. If the requirements are met, the agent retrieves the inputs from the knowledge graph, records the current timestamp, and changes the status of the derivation to `InProgress`. The inputs are then passed to the method provided by the developer and transformed into outputs. At this point, the status of the derivation will be changed to `Finished` and the newly derived outputs are temporarily connected to it. This status refers to a distinction made between a task that has been completed but still requires post-processing or cleaning-up, and a task that is complete in the sense that it requires no further action. It is used to prevent multiple agents from trying to perform the same cleaning-up tasks simultaneously and is removed when the derivation subgraph is tidied up during the next scheduled monitoring period. The cleaning-up process includes deleting the old instances, connecting the new instances with the original derivation and any downstream derivations that exist, removing all the status information, and finally updating the timestamp of the derivation to keep the derivation subgraph current. The monitoring is performed at a scheduled time interval and its frequency can be user-defined. If an error occurs during any operation, the agent changes the status of the derivation to `Error` and records the exception trace.

3.2.3. Concurrency and multi-threading

Being a decentralised system by design, the World Avatar contains many agents that are operating on the knowledge graph simultaneously. In situations where multiple agents request updates to the same derivation, the corresponding agent must handle concurrent requests correctly and efficiently. For example, such a situation arises when instance *i1* and *i2* illustrated in the generic form of DAG in Fig. 4 are accessed at the same time. In synchronous communication mode, the current implementation ensures that no duplicated information is added to the knowledge graph through the use of the SPARQL update detailed as Query 4 in Appendix A.3. In asynchronous communication mode, the agent uses the data property `uuidLock` to identify and lock the thread currently handling a derivation, avoiding duplicated execution. These measures ensure that concurrency is handled correctly without sacrificing the high throughput of multithreading, paving the way for distributed agent deployment.

3.3. Derivation client

Having established the ontology to capture the derivation process and the agent template to perform the derivation update, we introduce here the derivation client capable of managing the derivation subgraphs. This involves determining if a derivation is out-of-date and, if so, requesting an update. This section discusses three cases in which updates to the derivation subgraph are handled using different communication modes, namely: purely synchronous, purely asynchronous, and mixed-type. For each case, we first present the general algorithm and then provide examples to describe the intended outcome.

3.3.1. Purely synchronous update

Determining the timeliness of each derivation and performing the necessary updates in a derivation subgraph is a recursive process. Given any derivation instance where the accessed information is derived from, the framework treats it as root, traverses upstream all the way to the derivation that is derived from source information, *i.e.*, all inputs of whom are not derived from anything, and finally updates derivations backwards. This may be described as a depth-first search (DFS) algorithm, presented in Algorithm 1.

```
Algorithm 1: Update synchronous derivations
Input: IRI of the root derivation instance
Result: The root derivation and all its upstream derivations are
        updated if deemed outdated
Create an empty directed acyclic graph $G$;
Cache root derivation $d$ and all its upstream derivations
 recursively in $G$;
updateSyncDerivation($d$, $G$);
Function updateSyncDerivation($d$, $G$):
    $U \leftarrow d.upstreams()$ ;        /* get immediate upstream
     derivations */
    if vertex $d \notin G.vertices$ then
        Add $d$ as vertex in $G$;
    end
    for $u \in U$ do
        if vertex $u \notin G.vertices$ then
            $visited_u \leftarrow false$;
            Add $u$ as vertex in $G$;
        else
            $visited_u \leftarrow true$;
        end
        if edge $(d,u) \notin G.edges$ then
            $traversed_{d,u} \leftarrow false$;
            Add $(d,u)$ as edge in $G$ ;   /* will throw error if
             circular dependency detected */
        else
            $traversed_{d,u} \leftarrow true$;
        end
        if $visited_u == false$ and $traversed_{d,u} == false$ then
            updateSyncDerivation($u$, $G$);
        end
    end
    Fire request to update $d$ if it is deemed outdated;
    Update outputs of $d$ in cache;
```

Fig. 7 illustrates a notable detail of Algorithm 1 that uses the DAG G to track the traversing of the graph. To do this, the algorithm adds derivation instances as vertices and connections between them as directed edges to G. Depending on the root derivation, the DFS algorithm can result in two versions of G, leaving out parallel branches that do not require updating. The update is carried out only for the derivations in the resulting graph during the DFS algorithm's backtracking. It should also be noted that, the computation only proceeds when both node and edge are previously unseen. For example, regardless of which branch is traversed first (derivation *d1* or *d2*) in the upper resulting graph, derivation *d0* is only visited the first time when branching. This design ensures the relevant upstream information are only visited once to avoid duplication of work.

Fig. 8 illustrates the simplest form of derivation update, *i.e.*, updating one synchronous derivation. For demonstration, we take the simplified representation of derivation expressed in Fig. 3 at timestamp 36 as a starting point. As aforementioned, this derivation is deemed outdated when comparing its timestamp with that of its inputs. Assuming

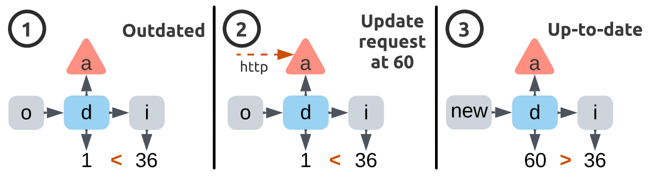


Fig. 8. The process of updating a single synchronous derivation.

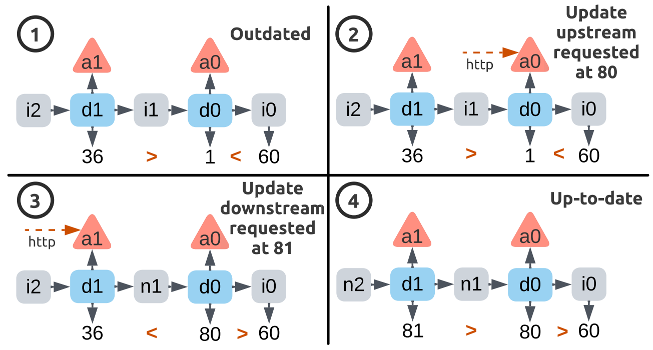

Outdated
Update upstream requested at 80
Update downstream requested at 81
Up-to-date
http

Fig. 9. The process of updating a derivation DAG consisting of only synchronous derivations.

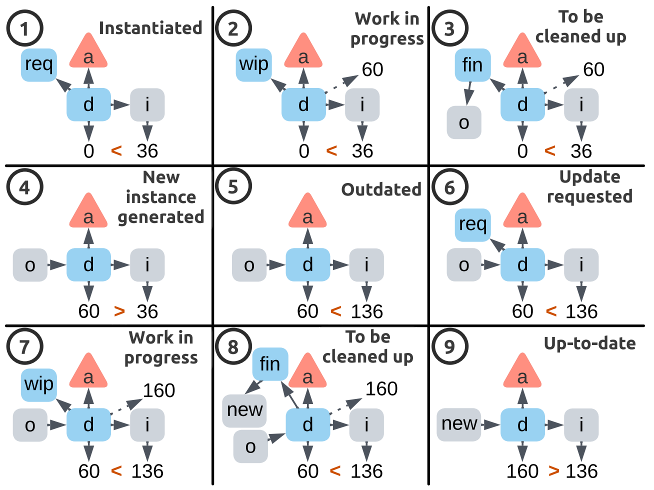


Fig. 10. The process of updating a single asynchronous derivation. The integers attached to the derivation instances via the dashed arrows denote the timestamps recorded by the data property `retrievedInputsAt`.

the output information is accessed at 60, the framework fires an update request to the agent associated with the derivation instance. Upon receiving the update request, the agent starts a calculation immediately and updates the output entity in the knowledge graph with the newly derived information. The derivation instance is thus up-to-date and the results are returned.

To expand this example to a slightly more complex situation, we consider accessing information *i2* in a linear chain of two synchronous derivations, as represented in Fig. 9. In this example, the input information of derivation *d1* belongs to derivation *d0*, therefore, the timeliness of *d1* is determined by comparing it with the derivation *d0*. However, just comparing the timestamps of these two will lead to an incorrect conclusion that *d1* is up-to-date. On the contrary, *d1* should be considered as outdated as it depends on an outdated derivation. Therefore, in the situation of assessing the timeliness of derivations that depend on derived information, the framework determines the timeliness of the upstream derivation (*d0*) first before performing any action to the downstream derivation (*d1*). For the presented example, assuming that information *i2* is accessed at time 80, the framework updates *d0* first, then *d1* immediately afterwards. With the new outputs connected in the knowledge graph, both derivations will be seen as up-to-date.

3.3.2. Purely asynchronous update

In the scientific computing domain for example, it is common to have situations where lengthy calculations from source input data are requested, requiring minutes or hours of wall time before the outputs are available. The asynchronous update suitable for such situations is discussed in this section, as described in Algorithm 2. It is similar to Algorithm 1, except that the algorithm for asynchronous derivations does not cache the derivation subgraph. Rather, the immediate upstream derivations are queried on-the-fly and hence their timeliness is determined purely based on real-time queries of the knowledge graph. The purpose of this design is to account for the fact that, due to the relatively long time scales involved, any information in the subgraph of asynchronous derivations can change while the process of carrying out an update is taking place.

Algorithm 2: Update asynchronous derivations

```
Input: IRI of the root derivation instance
Result: The root derivation and all its upstream derivations are
        requested for update if deemed outdated
Create an empty directed acyclic graph $G$;
Query derivation instance $d$ from the KG using the given
 rootDerivationIRI;
updateAsyncDerivation($d$, $G$);
Function updateAsyncDerivation($d$, $G$):
    Query the list $U$ of all immediate upstream derivations of $d$;
    if $vertex\ d \notin G.vertices$ then
        Add $d$ as vertex in $G$;
    end
    for $u \in U$ do
        if $vertex\ u \notin G.vertices$ then
            $visited_u \leftarrow false$;
            Add $u$ as vertex in $G$;
        else
            $visited_u \leftarrow true$;
        end
        if $edge\ (d,u) \notin G.edges$ then
            $traversed_{d,u} \leftarrow false$;
            Add $(d,u)$ as edge in $G$ ;   /* will throw error if
             circular dependency detected */
        else
            $traversed_{d,u} \leftarrow true$;
        end
        if $visited_u == false$ and $traversed_{d,u} == false$ then
            updateAsyncDerivation($u$, $G$);
        end
    end
    Mark $d$ as Requested if it is deemed outdated;
```

As depicted by Fig. 10, we start from the point in time when the derivation is just instantiated, *i.e.* the asynchronous derivation is initialised with the status as `Requested` and a timestamp of 0, with no output computed. The actual update of an asynchronous derivation will be dealt with by the derivation agent and is not concurrent with the

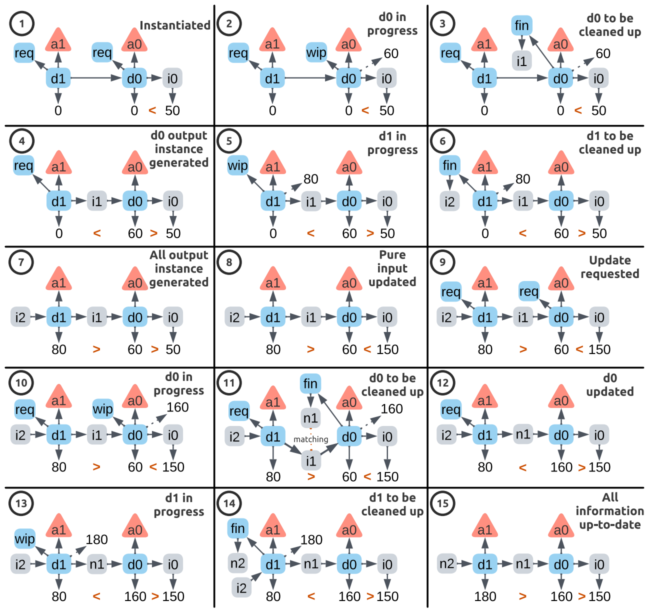


Fig. 11. The process of updating a derivation DAG consisting of only asynchronous derivations. The integers attached to the derivation instances via the dashed arrows denote the timestamps recorded by the data property `retrievedInputsAt`.

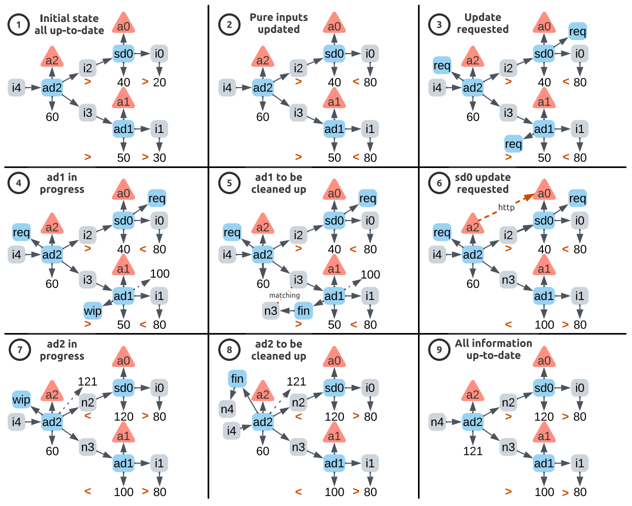


Fig. 12. The process of updating a derivation DAG consisting of asynchronous derivations that are dependent on synchronous derivations.

request for that update. As the agent periodically checks the status of derivations that are derived using itself, the requested derivation will be turned into `InProgress` at the next trigger and the timestamp when the inputs were read will be recorded. The successful completion of the job will be reflected in its status `Finished`. The agent will then connect the generated output to the derivation instance, update the timestamp and lastly remove the status altogether. The update process is similar to that of the initial computation. The only difference is that the agent will perform instance matching when reconnecting the newly derived information to existing downstream derivations in the derivation subgraph.

The same example can be expanded to a linear chain of two asynchronous derivations instantiated with only one piece of input data, as illustrated in Fig. 11. In this case, as none of the outputs are computed, the downstream derivation *d1* is directly marked as `isDerivedFrom` its upstream derivation *d0*. The agent responsible for *d0* will operate in the same way as aforementioned, the only difference being that the agent will reconnect the new derived instance *i1* as input to derivation *d1* and remove the direct connection between the two derivations. Similar to scientific workflow, the agent that manages the downstream derivation *d1* can begin its execution only after its predecessors have finished successfully. Therefore, block 1 to 7 of Fig. 11 showcases one desired usage of the derivation framework to automatically complete a predefined workflow given input data. Once all derived instances are computed, we illustrate the steps for updating the derivations in block 8 to 15, where the source input data is updated. Upon request, the algorithm traverses the derivation chain and determines the timeliness of the derivations. Unlike synchronous update, the algorithm only marks derivations as `Requested`, leaving the actual update to individual agents in the same way as aforementioned.

3.3.3. Mixed-type update

The final example we provide is a derivation subgraph consisting of mixed-type derivations. Specifically, asynchronous derivations depend on synchronous derivations, *i.e.*, the lengthy calculation relies on results from fast computations. It is worth noting that the other way would lead to a purely asynchronous scenario. If a synchronous derivation is dependent on an asynchronous one, the synchronous derivation effectively becomes asynchronous due to the need to wait a long time for the response. For that reason, the case of asynchronous derivations depending on synchronous ones is the only mixed case that needs to be considered, without loss of generality.

As previously shown, we need to determine if a derivation is out-of-date and perform/request an update. These two parts are combined and executed in a recursive algorithm for derivation subgraphs that only consist of synchronous derivations. However, to save computation resources for updating mixed-type derivation subgraphs, we would like to design the framework to work in a way that the update of upstream synchronous derivations is only computed when the agent updates the asynchronous derivation. Therefore, when the algorithm recursively determines the timeliness of the asynchronous derivations, markup is needed in the knowledge graph to indicate that the downstream asynchronous derivation is in fact out-of-date and, hence, should be marked as `Requested`.

A unified method is provided as a wrapper function of the two algorithms previously introduced. Depending on the instantiated type of root derivation, the function chooses a different entry algorithm. Fig. 12 demonstrates the lifecycle of such an example. Synchronous derivation *sd0* will be marked as `Requested` when the algorithm traverses, which serves as a signal when the algorithm determines the timeliness of downstream asynchronous derivation *ad2*. The agent monitoring *ad2* will request an update for *sd0* and start its execution after it confirmed that all its immediate upstream asynchronous derivations are up-to-date.

So far we have introduced three parts for the derived information framework that work together in a cohesive fashion: the derivation ontology, the derivation agent, and the derivation client. The fundamental objective of the framework is to ensure that all provenance information is explicitly documented within the knowledge graph, capturing details on what calculations were performed, their origins, and the functions (agents) employed. This allows for the temporal coherence, *i.e.* the ordering of the timestamps of events. As such its correctness can be verified by the provided function `validateDerivations`, which ensures there are no circular dependencies, and each instance (pure inputs and derivations) has a valid timestamp. Additionally, we have included test cases to cover the key features and functionalities of the framework. There also exist minimal working examples in both Java and Python as tutorials for newcomers. For these resources, please refer to TheWorldAvatar repository open-souced on GitHub[1] and PyPI[2]. We believe this transparency facilitates verification of the accuracy of the developed framework.

4. Use case

Flood impacts can be classified as direct versus indirect and tangible versus intangible effects [56]. This work serves as a proof-of-concept for the automatic assessment of flood impacts focusing on direct tangible impacts, referring to assets at risk due to direct physical exposure to floodwater. The intangible and indirect impacts of a potential flooding event are not considered in this work.

The impact estimation considers the number of buildings and the total property value that are potentially at risk by a potential flooding event. This involves information and events that occur at varying rates. However, existing approaches primarily focus on static evaluations and scenario planning [56], lacking the capability to offer a real-time view of the value at risk. This limitation highlights the necessity for cross-domain interoperability to adequately address the dynamic nature of such events. As demonstrated in the previous section, the derivation framework semantically annotates the dependencies between different pieces of information and uses computational agents to keep them current as a living snapshot of real-world events, making it an ideal candidate to address this problem.

The details of the domain ontologies and agent logic can be found in [57]. Here, we focus on how the derivation framework is used for this UK-based use case with a simplified example, including the coverage of the derivation subgraph and the processes that occur as information is cascaded over time. The results for the actual data are presented using the Digital Twin Visualisation Framework (DTVF) that is part of the World Avatar.

4.1. Automated population and update of the derivation subgraph

The flood impact assessment uses data from various sources, including application programming interfaces (APIs) such as the Environment Agency Real Time Flood-Monitoring API [58], Energy Performance Certificates (EPC) [59], and HM Land Registry Open Data [60]. These data are instantiated and updated in the World Avatar knowledge graph regularly by input agents. Using the source information, the impact assessment involves various derivation agents that work together to populate a derivation subgraph. This process encompasses two types of actions on the derivation subgraph: creating newly derived information, such as the impact of a newly issued flood warning, and updating existing information, such as the updated impact of an existing flood warning when some of the source information is updated.

One of these derivation agents, the Flood Assessment Agent, calculates the flood impact by identifying the buildings located within the affected area and determining the total market value of the properties at risk. The property value of each building is estimated by either scaling its most recent transaction record based on the local property price index or by multiplying its floor area by the average square metre price for its postal code. This representative average price can be computed by the Average Square Metre Price Agent considering all of the most recent transaction records in the area.

Fig. 13 illustrates progresses in the derivation subgraph when evaluating the potential impact of an issued flood warning. As denoted by the red dashed arrow, the flood warning instance is instantiated by the Flood Instantiation Agent. It specifies information about the expected severity of a flood event and the specific geospatial extent that is at risk. In this simplified example, the geospatial polygon is assumed to cover a postal code that includes two buildings, each with data about its

[1] https://github.com/cambridge-cares/TheWorldAvatar
[2] https://pypi.org/project/pyderivationagent/

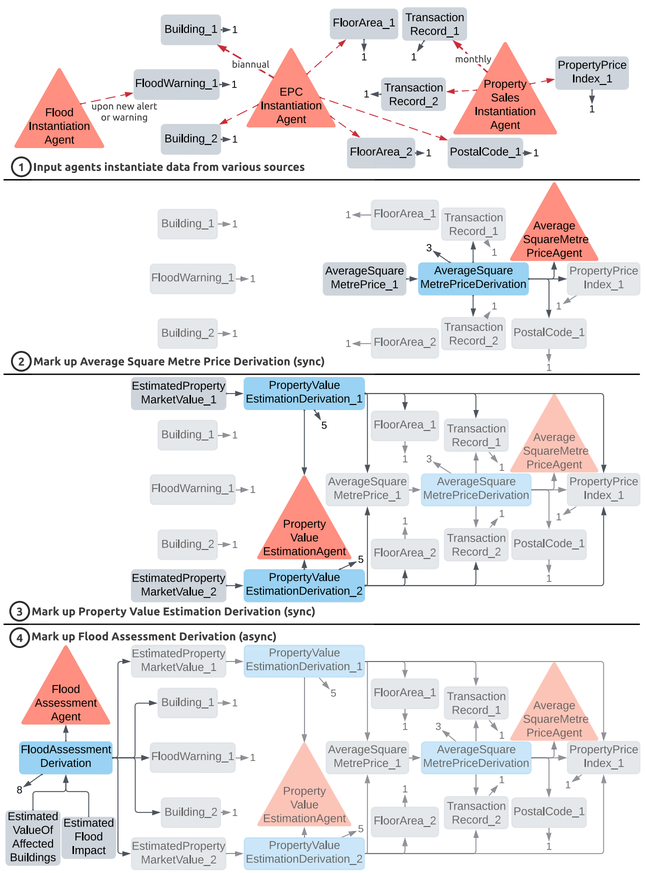
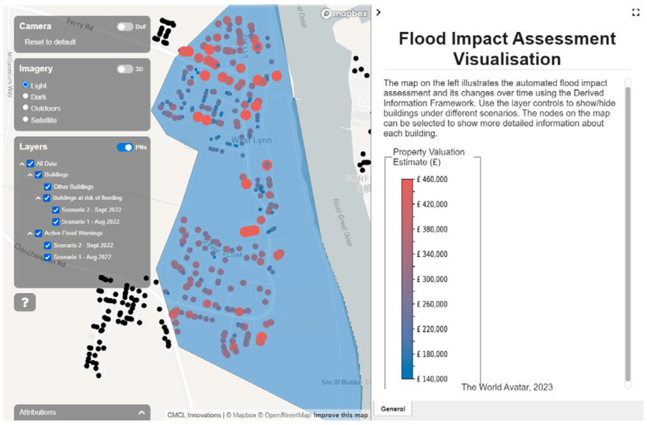

Building_1
FloorArea_1
Transaction Record_1
biannual
monthly
PropertyPrice Index_1
FloodWarning_1
EPC Instantiation Agent
Transaction Record_2
Property Sales Instantiation Agent
Flood Instantiation Agent
upon new alert or warning
Building_2
FloorArea_2
PostalCode_1
1 Input agents instantiate data from various sources
Average SquareMetre PriceAgent
AverageSquare MetrePrice_1
AverageSquare MetrePriceDerivation
2 Mark up Average Square Metre Price Derivation (sync)
EstimatedProperty MarketValue_1
PropertyValue EstimationDerivation_1
Property Value EstimationAgent
EstimatedProperty MarketValue_2
PropertyValue EstimationDerivation_2
3 Mark up Property Value Estimation Derivation (sync)
4 Mark up Flood Assessment Derivation (async)
Flood Assessment Agent
FloodAssessment Derivation
Estimated ValueOf Affected Buildings
Estimated Flood Impact

Fig. 13. The process of populating the derivation subgraph for flood impact assessment. Entities with low opacity represent existing entities in the knowledge graph, *i.e.* added in the previous steps of agents' operation before the agent adds newly derived information. The integers attached to the entities denote exemplary timestamps at which this information has been added to the knowledge graph.

floor area and its most recent transaction record. Additionally, the UK property price index which captures changes in the value of residential properties is instantiated as a time series in the knowledge graph and is updated on a monthly basis.

Once the derivation agents are deployed, the population of derived information starts with marking up the average square metre price derivation for all postcodes within the region of interest. Next, the property value estimation derivation is marked up for all buildings. Since these computations are relatively fast, they are marked up as synchronous derivations to obtain instantaneous responses. The flood assessment derivation is marked up upon the instantiation of the flood warning instance. In practice, a flood warning can cover more than a dozen postal codes with hundreds of buildings. Its computation can take some time and is thus marked up as an asynchronous derivation. Consequently, the absence of a status associated with the flood impact assessment derivation indicates that the derived information is up-to-date, signifying the completion of the assessment process. All derivation markups are instantiated programmatically through rigorously tested SPARQL and geospatial queries to ensure the accuracy of the captured dependencies.

As each input agent operates at different frequencies, the impact of a flood can alter when the source information is updated while the warning is still active. These markups are dynamically updated by the agents to integrate newly available information. For example, the property price indices for all administrative districts in the UK are updated monthly, and the geospatial extent of an active flood warning can also change, both of which can result in outdated flood impact

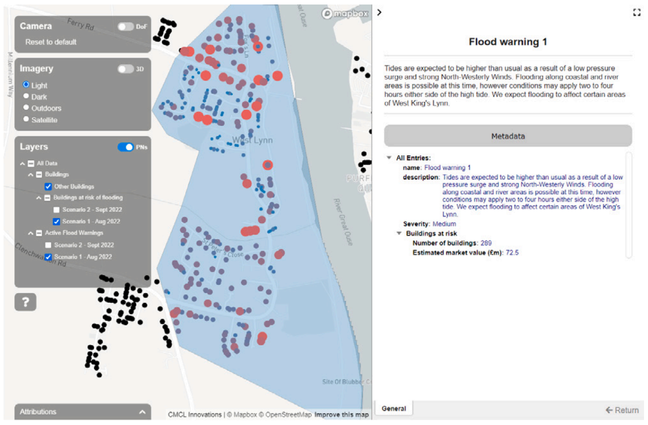


Flood Impact Assessment Visualisation

The map on the left illustrates the automated flood impact assessment and its changes over time using the Derived Information Framework. Use the layer controls to show/hide buildings under different scenarios. The nodes on the map can be selected to show more detailed information about each building.

The World Avatar, 2023

Fig. 14. Web page visualising a flood impact assessment with layers for buildings and flood warnings. The blue region denotes a potential flooding area. The colour and size of the dots in this region reflect estimated property values. Buildings outside of the region are considered unaffected and are thus marked with black dots.

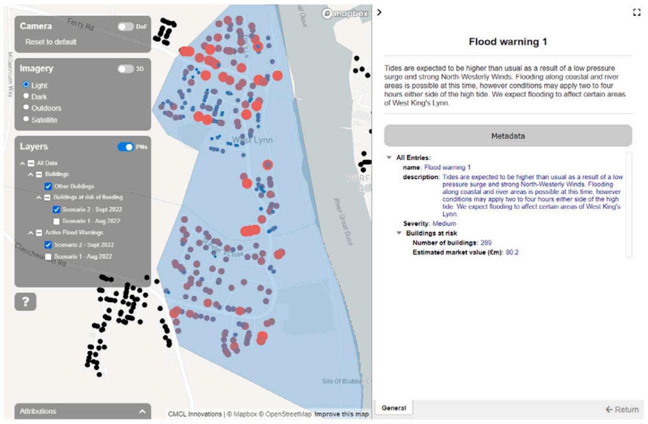


Flood warning 1

Tides are expected to be higher than usual as a result of a low pressure surge and strong North-Westerly Winds. Flooding along coastal and river areas is possible at this time, however conditions may apply two to four hours either side of the high tide. We expect flooding to affect certain areas of West King's Lynn.

All Entries:
name: Flood warning 1
description: Tides are expected to be higher than usual as a result of a low pressure surge and strong North-Westerly Winds. Flooding along coastal and river areas is possible at this time, however conditions may apply two to four hours either side of the high tide. We expect flooding to affect certain areas of West King's Lynn.
Severity: Medium
Buildings at risk
Number of buildings: 289
Estimated market value (£m): 72.5

(a) Scenario 1: Impact assessment of a newly issued flood warning.



Flood warning 1

Tides are expected to be higher than usual as a result of a low pressure surge and strong North-Westerly Winds. Flooding along coastal and river areas is possible at this time, however conditions may apply two to four hours either side of the high tide. We expect flooding to affect certain areas of West King's Lynn.

All Entries:
name: Flood warning 1
description: Tides are expected to be higher than usual as a result of a low pressure surge and strong North-Westerly Winds. Flooding along coastal and river areas is possible at this time, however conditions may apply two to four hours either side of the high tide. We expect flooding to affect certain areas of West King's Lynn.
Severity: Medium
Buildings at risk
Number of buildings: 289
Estimated market value (£m): 80.2

(b) Scenario 2: Impact update of an existing flood warning after property price index has increased.

Fig. 15. Automated flood impact assessment and update using the derived information framework. The side panel displays information about the clicked feature that has been dynamically retrieved from the knowledge graph.

assessments. Like the update of any other derived information, the impact assessment is designed to be on request to save computational power. Upon an update request, the derivation framework traverses the derivation subgraph to check if the average square metre price is up-to-date and if not, it backtracks to mark the derivations as outdated which triggers the flood assessment as "requested". Subsequently, Flood Assessment Agent triggers an update by calling the Average Square Metre Price Agent directly, *i.e.* synchronously. The price is updated in the knowledge graph which is further utilised by the Property Value Estimation Agent to update the estimated property market value. Only when all the input information is updated, the Flood Assessment Agent starts the impact estimate. This guarantees the information used in the assessment faithfully reflects the real world. The use case demonstrates both communication modes of the derived information framework by utilising both synchronous and asynchronous derivations.

4.2. *Visualisation of potential flood impact*

Following the simplified example, we present the visualisation of the impact assessment using real data. The derivation markup was created for a flood warning covering 34 postal codes and 289 buildings altogether. It is important to highlight that although there is no specific metric employed to directly assess whether all buildings within the flood area are accounted for by the derivation framework, the coverage of inputs can be evaluated by comparing the buildings identified as at risk with the polygon of the potential flooding area via SPARQL queries. To ensure the accuracy of results, all agents developed in this work are tested on synthetic data. For the real-world building information used in the study, data were obtained from EPC and HM Land Registry Open Data. However, the extent of completeness in their respective database lies beyond the scope of this work.

Fig. 14 visualises the estimated impact of a potential flood event. The representation separates different data into distinct layers, including buildings and the geospatial area affected by the active flood warning. By overlaying the flood boundary on the map, it is possible to identify which properties are at risk. The buildings within the flooded region are colour-coded with their estimated property values, while the other buildings in the administrative district are included only with their location information.

Fig. 15 presents two scenarios that are available to be selected on the plot: the estimated flood impact in August and September 2022. Between these two scenarios, an update of the district's property price index has been factored into calculating the total property value at risk. When clicking on the flooding area in different scenarios, the estimated impact and the detailed description of the flooding event are queried by the DTVF on-the-fly and displayed on the side panel. As the derivation agents manage the knowledge graph, the visualisation will be automatically updated.

5. Evaluation

The evaluation process in this study is conducted in three steps. Firstly, we assess the OntoDerivation TBox. Subsequently, we present the results obtained from scalability tests for the framework. Finally, we discuss interoperability between OntoDerivation and PROV-O.

5.1. *Ontology*

HermiT reasoner [61] and OntoDebug plugin [62] for the Protégé ontology editor [63] were used for this evaluation. The OntoDerivation implements 8 classes, 5 object properties and 2 data properties. The HermiT reasoner is able to classify the OntoDerivation ontology. In the debugging mode, OntoDebug detects the ontology as both *coherent* and *consistent*. Protégé does not show any errors related to the illegal declaration of entities or reuse of entities. All changes to OntoDerivation TBox are version-controlled in GitHub.

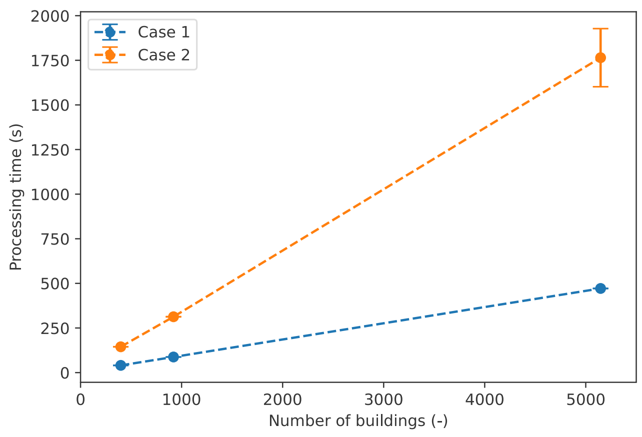


Fig. 16. Processing time when the derivation framework updates flood impact assessment for potential flooding events with different amounts of buildings at risk. Case 1: the instantiation of new flood warnings only. Case 2: the update of the property price index, followed by the instantiation of new flood warnings.

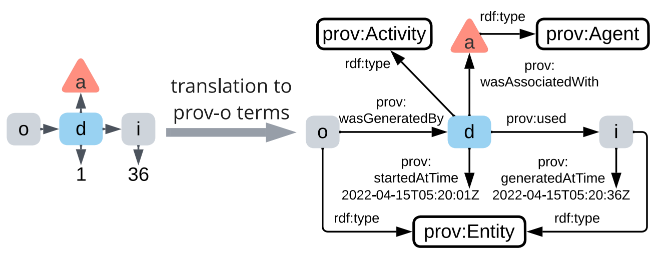


Fig. 17. An example translation of a derivation instance from OntoDerivation to PROV-O terms. The Unix timestamps are converted to `xsd:dateTime`.

5.2. Scalability

To examine the performance and scalability of the derivation framework in handling various scales of flooding events, tests were carried out on a virtual machine hosted on DigitalOcean. The virtual machine features 4 Intel Xeon Gold 6248 CPUs operating at 2.50 Ghz and 32 GB DIMM RAM. The triplestore and agents are deployed as individual docker containers. The analysis consists of two cases: (1) the instantiation of new flood warnings only, and (2) the update of the property price index, followed by the instantiation of new flood warnings. The second case triggers updates for the average square metre price and estimated property market value of all buildings potentially at risk. Tests for both cases were performed for three flood warnings each covering a different number of buildings at risk: 398, 920 and 5147. We measure from the time the framework picks up the derivation for flood impact (re-)assessment to the timestamp when all information is up-to-date. Each test was repeated at least three times to obtain the mean and standard deviations in processing time.

Fig. 16 illustrates the linear scalability of the derivation framework across the varying number of buildings examined. It should be noted, however, that the actual amount of processing time is use-case-specific. By deducing the processing time of case 1 from case 2, it is observed that the agents process approximately four buildings' derivations per second. At present, only one instance for each agent is deployed. In order to evaluate the performance gains achievable through the deployment of multiple instances of a given agent, the introduction of a load balancer becomes necessary which is however beyond the scope of this work.

5.3. Interoperability

As aforementioned, PROV-O serves as a generic ontology for representing provenance records, abstracting diverse needs within various domains.

Fig. 17 illustrates a practical implementation to enable interoperability between OntoDerivation and PROV-O. Given a derivation, we provide SPARQL Query 5 in Appendix A.3 that programmatically translates the derivation graph of the given derivation and all its upstream derivations into corresponding PROV-O terms. The resulting triples can be exported as linked data for publication and may be queried and processed using provenance tools.[3]

One aspect of interoperability that remains unexplored is the capability to manage (query & update) provenance records formulated in native PROV-O expressions. This pertains to the absence of status information in the exported provenance records due to the lack of inherent support for status details in PROV-O. The capability to possess this falls however beyond the scope of this work.

6. Conclusions

In this work, we developed a derived information framework as a knowledge-graph-native solution for tracking provenance and managing information within dynamic knowledge graphs. It expands previous capabilities and further abstracts complexity away from the developers of individual agents. The architecture includes a lightweight ontology that marks up agent communication as provenance records in the knowledge graph, an agent template that standardises the operation of agents in both synchronous and asynchronous communication modes, as well as a client library that offers functions for managing the derivation subgraph. The framework is technology-agnostic and is made available in both Java and Python.

To showcase the accessibility of the framework, it was applied to a flood impact assessment use case within the World Avatar project. The use case involves several derivation agents developed using the agent template. Once the source information is gathered by input agents from different APIs, the derivation subgraph is populated by creating derivation markups as requests for derivation agents to generate the derived information required in the impact assessment. If the input information is refreshed, the framework automatically updates the derived information when accessed. The results were visualised and can be deployed as a regular service if needed.

Future work includes expanding the framework to physical experimentation, implementing automated fault recovery for computations, evaluating the performance of different agents that can perform the same task, and incorporating automated service discovery for derivation markup generation.

CRediT authorship contribution statement

Jiaru Bai: Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Kok Foong Lee:** Conceptualization, Methodology, Software, Validation, Writing – review & editing. **Markus Hofmeister:** Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – review & editing, Visualization. **Sebastian Mosbach:** Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Jethro Akroyd:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Markus Kraft:** Conceptualization, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

[3] https://www.software.ac.uk/who-do-we-work/provenance-tool-suite

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This research was supported by the National Research Foundation, Singapore, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. Part of this work was supported by Towards Turing 2.0 under the EPSRC Grant EP/W037211/1 & The Alan Turing Institute. J. Bai acknowledges financial support provided by CSC Cambridge International Scholarship from Cambridge Trust and China Scholarship Council. M. Hofmeister acknowledges financial support provided by the Cambridge Trust and CMCL. M. Kraft gratefully acknowledges the support of the Alexander von Humboldt Foundation, Germany. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Appendix

A.1. Namespaces

The ontological representation in this work involves several namespaces as listed below.

om: <http://www.ontology-of-units-of-measure.org/resource/om-2/>
owl: <http://www.w3.org/2002/07/owl#>
rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
rdfs: <http://www.w3.org/2000/01/rdf-schema#>
xsd: <http://www.w3.org/2001/XMLSchema#>
time: <http://www.w3.org/2006/time#>
prov: <http://www.w3.org/ns/prov#>
OntoDerivation: <https://www.theworldavatar.com/kg/ontoderivation/>
OntoAgent: <http://www.theworldavatar.com/ontology/ontoagent/MSM.owl#>

A.2. Description logic representation of OntoDerivation

The description logic representations of the OntoDerivation developed in this work are provided below.

Classes:
Derivation $\sqsubseteq \top$
DerivationWithTimeSeries $\sqsubseteq \top$
DerivationAsyn $\sqsubseteq \top$
Status $\sqsubseteq \top$
Requested $\sqsubseteq$ Status
InProgress $\sqsubseteq$ Status
Finished $\sqsubseteq$ Status
Error $\sqsubseteq$ Status
Object Properties:
$\top \sqsubseteq \forall$ isDerivedFrom.$\top$
$\exists$ isDerivedFrom.$\top \sqsubseteq$ (Derivation $\sqcup$ DerivationAsyn $\sqcup$ DerivationWithTimeSeries)
$\top \sqsubseteq \forall$ isDerivedUsing.OntoAgent:Service
$\exists$ isDerivedUsing.$\top \sqsubseteq$ (Derivation $\sqcup$ DerivationAsyn $\sqcup$ DerivationWithTimeSeries)
$\exists$ belongsTo.$\top \sqsubseteq \top$
$\top \sqsubseteq \forall$ belongsTo.(Derivation $\sqcup$ DerivationAsyn $\sqcup$ DerivationWithTimeSeries)
$\top \sqsubseteq \forall$ hasStatus.Status
$\exists$ hasStatus.$\top \sqsubseteq$ DerivationAsyn
$\top \sqsubseteq \forall$ hasNewDerivedIRI.$\top$
$\exists$ hasNewDerivedIRI.$\top \sqsubseteq$ Finished
Data Properties:
$\exists$ retrievedInputsAt.$\top \sqsubseteq$ DerivationAsyn
$\top \sqsubseteq \forall$ retrievedInputsAt.xsd:decimal
$\exists$ uuidLock.$\top \sqsubseteq$ DerivationAsyn
$\top \sqsubseteq \forall$ uuidLock.xsd:string

A.3. Example queries

Listing 1: SPARQL query to obtain all derivation instances in the knowledge graph.

```
PREFIX OntoDerivation: <https://www.
    theworldavatar.com/kg/ontoderivation/>
PREFIX time: <http://www.w3.org/2006/time#>

SELECT ?derivation ?devTime ?inputTime ?status ?
    status_type
WHERE {
  VALUES ?derivationType {
    OntoDerivation:DerivationAsyn
    OntoDerivation:Derivation
    OntoDerivation:DerivationWithTimeSeries
  }
  ?derivation a ?derivationType ;
  time:hasTime/time:inTimePosition/time:
      numericPosition ?devTime .
  OPTIONAL {
    ?derivation OntoDerivation:isDerivedFrom ?
        upstream .
    ?upstream time:hasTime/time:inTimePosition/
        time:numericPosition ?inputTime .
  }
  OPTIONAL {
    ?derivation OntoDerivation:hasStatus ?status
        .
    ?status a ?status_type .
  }
}
```

Listing 2: SPARQL query to map the derivation inputs to agent I/O signature.

```
PREFIX OntoDerivation: <https://www.
    theworldavatar.com/kg/ontoderivation/>
PREFIX OntoAgent: <http://www.theworldavatar.com/
    ontology/ontoagent/MSM.owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-
    schema#>

SELECT DISTINCT ?input ?type
WHERE {
  <agentIRI> OntoAgent:hasOperation/OntoAgent:
      hasInput/
    OntoAgent:hasMandatoryPart/OntoAgent:hasType
        ?type .
  <derivationIRI> OntoDerivation:isDerivedFrom ?
      input .
  ?input a*/rdfs:subClassOf* ?type .
}
```

Listing 3: SPARQL query to determine the immediate upstream derivations that requires an update.

```
PREFIX OntoDerivation: <https://www.
    theworldavatar.com/kg/ontoderivation/>
PREFIX time: <http://www.w3.org/2006/time#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT DISTINCT ?upstreamDerivation ?
    upstreamDerivationType
WHERE {
  {
    <derivationIRI> OntoDerivation:isDerivedFrom/
      OntoDerivation:belongsTo? ?
          upstreamDerivation .
    ?upstreamDerivation a ?upstreamDerivationType
          .
    VALUES ?upstreamDerivationType {
      OntoDerivation:Derivation
      OntoDerivation:DerivationWithTimeSeries
      OntoDerivation:DerivationAsyn
    }
  }
  ?upstreamDerivation time:hasTime/time:
      inTimePosition/
    time:numericPosition ?
        upstreamDerivationTimestamp .
  OPTIONAL {
    ?upstreamDerivation OntoDerivation:hasStatus
        ?status .
    {
      ?status a ?statusType .
      FILTER (?statusType != owl:Thing && ?
          statusType != owl:NamedIndividual)
     }
  }
  OPTIONAL {
    ?upstreamDerivation OntoDerivation:
        isDerivedFrom/time:hasTime/
      time:inTimePosition/time:numericPosition ?
          pureInputTimestamp .
  }
  OPTIONAL {
    ?upstreamDerivation OntoDerivation:
        isDerivedFrom/OntoDerivation:belongsTo/
      time:hasTime/time:inTimePosition/time:
          numericPosition
      ?inputsBelongingToDerivationTimestamp .
  }
  FILTER (?upstreamDerivationTimestamp < ?
      pureInputTimestamp ||
    ?upstreamDerivationTimestamp < ?
        inputsBelongingToDerivationTimestamp ||
    ?statusType = OntoDerivation:Requested ||
    ?statusType = OntoDerivation:InProgress ||
    ?statusType = OntoDerivation:Finished ||
    ?statusType = OntoDerivation:Error)
}
```

Listing 4: SPARQL query to update a given derivation in the knowledge graph.

```
PREFIX OntoDerivation: <https://www.
    theworldavatar.com/kg/ontoderivation/>
PREFIX time: <http://www.w3.org/2006/time#>

DELETE {
  ?e ?p1 ?o .
  ?s ?p2 ?e .
  ?d OntoDerivation:hasStatus ?status .
  ?status a ?statusType .
  ?timeIRI time:numericPosition ?dTs .
}
INSERT {
  <newInstance1> a <rdfTypeOfNewInstance1> .
  <newInstance2> a <rdfTypeOfNewInstance2> .
  <newInstance3> a <rdfTypeOfNewInstance3> .
  <newInstance3> OntoDerivation:belongsTo <
      derivationIRI> .
  <newInstance1> OntoDerivation:belongsTo <
      derivationIRI> .
  <downstreamDerivation1> OntoDerivation:
      isDerivedFrom <newInstance1> .
  <downstreamDerivation2> OntoDerivation:
      isDerivedFrom <newInstance1> .
  <newInstance2> OntoDerivation:belongsTo <
      derivationIRI> .
  <downstreamDerivation3> OntoDerivation:
      isDerivedFrom <newInstance2> .
  ?timeIRI time:numericPosition 1659981343 .
}
WHERE {
  {
    SELECT ?d ?timeIRI ?dTs ?status ?statusType ?
        e ?p1 ?o ?s ?p2
    WHERE {
      {
        VALUES ?d { <derivationIRI> }
        ?d time:hasTime/time:inTimePosition ?
            timeIRI .
        ?timeIRI time:numericPosition ?dTs .
        ?d OntoDerivation:isDerivedFrom/
            OntoDerivation:belongsTo? ?ups .
        ?ups time:hasTime/time:inTimePosition/
            time:numericPosition ?upsTs .
        FILTER (?dTs < ?upsTs)
      }
      {
        ?e OntoDerivation:belongsTo ?d .
        ?e ?p1 ?o .
        OPTIONAL { ?s ?p2 ?e . }
      }
      OPTIONAL {
        ?d OntoDerivation:hasStatus ?status .
        ?status a ?statusType .
      }
    }
  }
}
```

Listing 5: SPARQL query to construct provenance records for a given derivation instance and all its upstream derivations using PROV-O terms.

```
PREFIX OntoDerivation: <https://www.
    theworldavatar.com/kg/ontoderivation/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX time: <http://www.w3.org/2006/time#>
PREFIX prov: <http://www.w3.org/ns/prov#>
CONSTRUCT {
  ?agent a prov:Agent .
  ?derivation a prov:Activity .
  ?input a prov:Entity .
  ?output a prov:Entity .
  ?derivation prov:wasAssociatedWith ?agent .
  ?derivation prov:used ?input .
  ?output prov:wasGeneratedBy ?derivation .
  ?derivation prov:startedAtTime ?devDateTime .
  ?input prov:generatedAtTime ?inputDateTime .
}
WHERE {
  <derivationIRI> (OntoDerivation:isDerivedFrom/
      OntoDerivation:belongsTo)* ?derivation .
  ?derivation OntoDerivation:isDerivedUsing ?
      agent .
  ?derivation time:hasTime/time:inTimePosition/
      time:numericPosition ?devTime .
  BIND("1970-01-01T00:00:00Z"^^xsd:dateTime +
      STRDT(CONCAT("PT", STR(?devTime), "S"), xsd
      :duration) AS ?devDateTime)
  ?derivation OntoDerivation:isDerivedFrom ?input
      .
  ?output OntoDerivation:belongsTo ?derivation .
  OPTIONAL {
    ?input time:hasTime/time:inTimePosition/time:
        numericPosition ?inputTime .
    BIND("1970-01-01T00:00:00Z"^^xsd:dateTime +
        STRDT(CONCAT("PT", STR(?inputTime), "S"),
         xsd:duration) AS ?inputDateTime)
  }
}
```

References

[1] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, J. Taylor, Industry-scale knowledge graphs: Lessons and challenges, Commun. ACM 62 (8) (2019) 36–43, http://dx.doi.org/10.1145/3331166.

[2] P. Hitzler, A review of the semantic web field, Commun. ACM 64 (2) (2021) 76–83, http://dx.doi.org/10.1145/3397512.

[3] C. Gutierrez, J.F. Sequeda, Knowledge graphs, Commun. ACM 64 (3) (2021) 96–104, http://dx.doi.org/10.1145/3418294.

[4] A. Hogan, E. Blomqvist, M. Cochez, C. D'Amato, G.D. Melo, C. Gutierrez, S. Kirrane, J.E.L. Gayo, R. Navigli, S. Neumaier, A.-C.N. Ngomo, A. Polleres, S.M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, ACM Comput. Surv. 54 (4) (2022) 1–37, http://dx.doi.org/10.1145/3447772.

[5] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P.N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, C. Bizer, DBpedia – A large-scale, multilingual knowledge base extracted from wikipedia, Semant. Web 6 (2) (2015) 167–195, http://dx.doi.org/10.3233/SW-140134.

[6] T. Pellissier Tanon, D. Vrandečić, S. Schaffert, T. Steiner, L. Pintscher, From freebase to wikidata: The great migration, in: Proceedings of the 25th International Conference on World Wide Web, 2016, pp. 1419–1428, http://dx.doi.org/10.1145/2872427.2874809.

[7] J. Akroyd, S. Mosbach, A. Bhave, M. Kraft, Universal digital twin – a dynamic knowledge graph, Data-Centr. Eng. 2 (2021) e14, http://dx.doi.org/10.1017/dce.2021.10.

[8] J. Bai, L. Cao, S. Mosbach, J. Akroyd, A.A. Lapkin, M. Kraft, From platform to knowledge graph: Evolution of laboratory automation, JACS Au 2 (2) (2022) 292–309, http://dx.doi.org/10.1021/jacsau.1c00438.

[9] J. Hendler, Agents and the semantic web, IEEE Intell. Syst. 16 (2) (2001) 30–37, http://dx.doi.org/10.1109/5254.920597.

[10] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web, Sci. Am. 284 (5) (2001) 34–43, http://dx.doi.org/10.1038/scientificamerican0501-34, URL https://www.jstor.org/stable/10.2307/26059207.

[11] J. Zhao, C. Bizer, Y. Gil, P. Missier, S. Sahoo, Provenance requirements for the next version of RDF, 2010, W3C Workshop – RDF Next Steps, Stanford, CA, USA, June 26-27, URL https://www.w3.org/2009/12/rdf-ws/papers/ws08.

[12] P. Ciccarese, S. Soiland-Reyes, K. Belhajjame, A.J.G. Gray, C. Goble, T. Clark, PAV ontology: Provenance, authoring and versioning, J. Biomed. Semant. 4 (1) (2013) 37, http://dx.doi.org/10.1186/2041-1480-4-37, URL https://pav-ontology.github.io/pav/.

[13] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, J. Zhao, PROV-O: The PROV ontology, 2013, W3C Recommendation, URL https://www.w3.org/TR/prov-o/.

[14] DCMI Usage Board, DCMI metadata terms, 2020, URL https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

[15] DCMI Usage Board, Bibliographic ontology (BIBO) in RDF, 2016, URL https://www.dublincore.org/specifications/bibo/bibo/.

[16] L. Moreau, L. Ding, J. Futrelle, D.G. Verdejo, P. Groth, M. Jewell, S. Miles, P. Missier, J. Pan, J. Zhao, Open provenance model (OPM) OWL specification, 2010, URL https://openprovenance.org/opm/model/opmo.

[17] L.F. Sikos, D. Philp, Provenance-aware knowledge representation: A survey of data models and contextualized knowledge graphs, Data Sci. Eng. 5 (3) (2020) 293–316, http://dx.doi.org/10.1007/s41019-020-00118-0.

[18] J. Yu, R. Buyya, A taxonomy of workflow management systems for grid computing, J. Grid Comput. 3 (3) (2005) 171–200, http://dx.doi.org/10.1007/s10723-005-9010-8.

[19] E. Deelman, D. Gannon, M. Shields, I. Taylor, Workflows and e-science: An overview of workflow system features and capabilities, Future Gener. Comput. Syst. 25 (5) (2009) 528–540, http://dx.doi.org/10.1016/j.future.2008.06.012.

[20] J. Liu, E. Pacitti, P. Valduriez, M. Mattoso, A survey of data-intensive scientific workflow management, J. Grid Comput. 13 (4) (2015) 457–493, http://dx.doi.org/10.1007/s10723-015-9329-8.

[21] E. Deelman, G. Singh, M.-H. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G.B. Berriman, J. Good, A. Laity, J.C. Jacob, D.S. Katz, Pegasus: A framework for mapping complex scientific workflows onto distributed systems, Sci. Program. 13 (3) (2005) 219–237, http://dx.doi.org/10.1155/2005/128026.

[22] E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P.J. Maechling, R. Mayani, W. Chen, R.F. Da Silva, M. Livny, K. Wenger, Pegasus, a workflow management system for science automation, Future Gener. Comput. Syst. 46 (2015) 17–35, http://dx.doi.org/10.1016/j.future.2014.10.008.

[23] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludascher, S. Mock, Kepler: An extensible system for design and execution of scientific workflows, in: Proceedings of 16th International Conference on Scientific and Statistical Database Management, 2004, IEEE, 2004, pp. 423–424, http://dx.doi.org/10.1109/SSDM.2004.1311241.

[24] The Apache Software Foundation, Apache airflow, 2022, URL https://airflow.apache.org/, (Accessed 17 July 2022).

[25] P. Di Tommaso, M. Chatzou, E.W. Floden, P.P. Barja, E. Palumbo, C. Notredame, Nextflow enables reproducible computational workflows, Nat. Biotechnol. 35 (4) (2017) 316–319, http://dx.doi.org/10.1038/nbt.3820.

[26] W. Falcon, The PyTorch Lightning team, PyTorch lightning, 2019, http://dx.doi.org/10.5281/zenodo.3828935, URL https://github.com/Lightning-AI/lightning, (Accessed 14 July 2022).

[27] E. Lyons, G. Papadimitriou, C. Wang, K. Thareja, P. Ruth, J. Villalobos, I. Rodero, E. Deelman, M. Zink, A. Mandal, Toward a dynamic network-centric distributed cloud platform for scientific workflows: A case study for adaptive weather sensing, in: 2019 15th International Conference on EScience (EScience), IEEE, 2019, pp. 67–76, http://dx.doi.org/10.1109/eScience.2019.00015.

[28] E. Deelman, T. Peterka, I. Altintas, C.D. Carothers, K.K. van Dam, K. Moreland, M. Parashar, L. Ramakrishnan, M. Taufer, J. Vetter, The future of scientific workflows, Int. J. High Perform. Comput. Appl. 32 (1) (2018) 159–175, http://dx.doi.org/10.1177/1094342017704893.

[29] R.F. da Silva, H. Casanova, K. Chard, D. Laney, D. Ahn, S. Jha, C. Goble, L. Ramakrishnan, L. Peterson, B. Enders, D. Thain, I. Altintas, Y. Babuji, R.M. Badia, V. Bonazzi, T. Coleman, M. Crusoe, E. Deelman, F.D. Natale, P.D. Tommaso, T. Fahringer, R. Filgueira, G. Fursin, A. Ganose, B. Gruning, D.S. Katz, O. Kuchar, A. Kupresanin, B. Ludascher, K. Maheshwari, M. Mattoso, K. Mehta, T. Munson, J. Ozik, T. Peterka, L. Pottier, T. Randles, S. Soiland-Reyes, B. Tovar, M. Turilli, T. Uram, K. Vahi, M. Wilde, M. Wolf, J. Wozniak, Workflows community summit: Bringing the scientific workflows community together, 2021, URL https://arxiv.org/abs/2103.09181, (Accessed 18 July 2022).

[30] N. Dragoni, S. Giallorenzo, A.L. Lafuente, M. Mazzara, F. Montesi, R. Mustafin, L. Safina, Microservices: Yesterday, today, and tomorrow, in: Present and Ulterior Software Engineering, Springer, 2017, pp. 195–216, http://dx.doi.org/10.1007/978-3-319-67425-4_12.

[31] Dapr Authors, APIs for building portable and reliable microservices, 2022, URL https://dapr.io/, (Accessed 14 July 2022).

[32] F. Farazi, N.B. Krdzavac, J. Akroyd, S. Mosbach, A. Menon, D. Nurkowski, M. Kraft, Linking reaction mechanisms and quantum chemistry: An ontological approach, Comput. Chem. Eng. 137 (2020) 106813, http://dx.doi.org/10.1016/j.compchemeng.2020.106813.

[33] F. Farazi, J. Akroyd, S. Mosbach, P. Buerger, D. Nurkowski, M. Salamanca, M. Kraft, OntoKin: An ontology for chemical kinetic reaction mechanisms, J. Chem. Inf. Model. 60 (1) (2020) 108–120, http://dx.doi.org/10.1021/acs.jcim.9b00960.

[34] A. Chadzynski, N. Krdzavac, F. Farazi, M.Q. Lim, S. Li, A. Grisiute, P. Herthogs, A. von Richthofen, S. Cairns, M. Kraft, Semantic 3D city database - an enabler for a dynamic geospatial knowledge graph, Energy and AI 6 (2021) 100106, http://dx.doi.org/10.1016/j.egyai.2021.100106.

[35] N. Mohammadi, J.E. Taylor, Thinking fast and slow in disaster decision-making with smart city digital twins, Nat. Comput. Sci. 1 (12) (2021) 771–773, http://dx.doi.org/10.1038/s43588-021-00174-0.

[36] D. Garijo, Y. Gil, Augmenting PROV with plans in P-PLAN: Scientific processes as linked data, in: Proceedings of the 2nd International Workshop on Linked Science, Vol. 951, CEUR Workshop Proceedings, 2012, URL https://oa.upm.es/19478/, (Accessed 19 July 2023).

[37] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan, J.V. den Bussche, The open provenance model core specification (v1.1), Future Gener. Comput. Syst. 27 (6) (2011) 743–756, http://dx.doi.org/10.1016/j.future.2010.07.005.

[38] D. Garijo, Y. Gil, O. Corcho, Abstract, link, publish, exploit: An end to end framework for workflow sharing, Future Gener. Comput. Syst. 75 (2017) 271–283, http://dx.doi.org/10.1016/j.future.2017.01.008.

[39] L. Moreau, B. Ludäscher, I. Altintas, R.S. Barga, S. Bowers, S. Callahan, G. Chin, B. Clifford, S. Cohen, S. Cohen-Boulakia, S. Davidson, E. Deelman, L. Digiampietri, I. Foster, J. Freire, J. Frew, J. Futrelle, T. Gibson, Y. Gil, C. Goble, J. Golbeck, P. Groth, D.A. Holland, S. Jiang, J. Kim, D. Koop, A. Krenek, T. McPhillips, G. Mehta, S. Miles, D. Metzger, S. Munroe, J. Myers, B. Plale, N. Podhorszki, V. Ratnakar, E. Santos, C. Scheidegger, K. Schuchardt, M. Seltzer, Y.L. Simmhan, C. Silva, P. Slaughter, E. Stephan, R. Stevens, D. Turi, H. Vo, M. Wilde, J. Zhao, Y. Zhao, Special issue: The first provenance challenge, Concurr. Comput.-Pract. Exp. 20 (5) (2008) 409–418, http://dx.doi.org/10.1002/cpe.1233.

[40] R. Chard, J. Pruyne, K. McKee, J. Bryan, B. Raumann, R. Ananthakrishnan, K. Chard, I.T. Foster, Globus automation services: Research process automation across the space–time continuum, Future Gener. Comput. Syst. 142 (2023) 393–409, http://dx.doi.org/10.1016/j.future.2023.01.010.

[41] J. Akroyd, Z. Harper, D. Soutar, F. Farazi, A. Bhave, S. Mosbach, M. Kraft, Universal digital twin: Land use, Data-Centr. Eng. 3 (2022) http://dx.doi.org/10.1017/dce.2021.21.

[42] M. Hofmeister, S. Mosbach, J. Hammacher, M. Blum, G. Röhrig, C. Dörr, V. Flegel, A. Bhave, M. Kraft, Resource-optimised generation dispatch strategy for district heating systems using dynamic hierarchical optimisation, Appl. Energy 305 (2022) 117877, http://dx.doi.org/10.1016/j.apenergy.2021.117877.

[43] J. Akroyd, A. Bhave, G. Brownbridge, E. Christou, M. Hillman, M. Hofmeister, M. Kraft, J. Lai, K.F. Lee, S. Mosbach, D. Nurkowski, O. Parry, CReDo technical paper 1: Building a cross-sector digital twin, 2022, URL https://doi.org/10.17863/CAM.81779, (Accessed 28 October 2023).

[44] T. Savage, J. Akroyd, S. Mosbach, M. Hillman, F. Sielker, M. Kraft, Universal digital twin–the impact of heat pumps on social inequality, Adv. Appl. Energy 5 (2022) 100079, http://dx.doi.org/10.1016/j.adapen.2021.100079.

[45] X. Zhou, A. Eibeck, M.Q. Lim, N.B. Krdzavac, M. Kraft, An agent composition framework for the J-park simulator - a knowledge graph for the process industry, Comput. Chem. Eng. 130 (2019) 106577, http://dx.doi.org/10.1016/j.compchemeng.2019.106577.

[46] S. Mosbach, A. Menon, F. Farazi, N. Krdzavac, X. Zhou, J. Akroyd, M. Kraft, Multiscale cross-domain thermochemical knowledge-graph, J. Chem. Inf. Model. 60 (12) (2020) 6155–6166, http://dx.doi.org/10.1021/acs.jcim.0c01145.

[47] J. Bai, R. Geeson, F. Farazi, S. Mosbach, J. Akroyd, E.J. Bringley, M. Kraft, Automated calibration of a poly(oxymethylene) dimethyl ether oxidation mechanism using the knowledge graph technology, J. Chem. Inf. Model. 61 (4) (2021) 1701–1717, http://dx.doi.org/10.1021/acs.jcim.0c01322.

[48] N. Lopes, A. Zimmermann, A. Hogan, G. Lukácsy, A. Polleres, U. Straccia, S. Decker, RDF needs annotations, 2010, W3C Workshop – RDF Next Steps, Stanford, CA, USA, June 26-27, URL https://www.w3.org/2009/12/rdf-ws/papers/ws09.

[49] O. Hartig, B. Thompson, Foundations of an alternative approach to reification in RDF, 2021, URL https://arxiv.org/abs/1406.3399v3.

[50] World Wide Web Consortium (W3C), RDF-star, 2021, URL https://w3c.github.io/rdf-star/.

[51] blazegraph, Reification done right, 2020, URL https://github.com/blazegraph/database/wiki/Reification_Done_Right.

[52] ontotext, What is RDF-star?, 2022, URL https://www.ontotext.com/knowledgehub/fundamentals/what-is-rdf-star/.

[53] J. Bai, K.F. Lee, S. Mosbach, 2023, URL https://github.com/cambridge-cares/TheWorldAvatar/tree/main/JPS_Ontology/ontology/ontoderivation, (Accessed 6 Feb 2023).

[54] S. Cox, C. Little, J.R. Hobbs, F. Pan, Time ontology in OWL, 2020, W3C Candidate Recommendation 26 March 2020, URL https://www.w3.org/TR/owl-time/, (Accessed 21 July 2022).

[55] M. Krämer, H.M. Würz, C. Altenhofen, Executing cyclic scientific workflows in the cloud, J. Cloud Comput. 10 (1) (2021) 1–26, http://dx.doi.org/10.1186/s13677-021-00229-7.

[56] M. Hammond, A. Chen, S. Djordjević, D. Butler, O. Mark, Urban flood impact assessment: A state-of-the-art review, Urban Water J. 12 (1) (2013) 14–29, http://dx.doi.org/10.1080/1573062x.2013.857421.

[57] M. Hofmeister, J. Bai, G. Brownbridge, S. Mosbach, K.F. Lee, F. Farazi, M. Hillman, M. Agarwal, S. Ganguly, J. Akroyd, M. Kraft, Semantic agent framework for automated flood assessment using dynamic knowledge graphs, 2023, (under review). Preprint available from https://como.ceb.cam.ac.uk/preprints/309/.

[58] Department for Environment Food & Rural Affairs, Real Time flood-monitoring API, 2021, URL https://environment.data.gov.uk/flood-monitoring/doc/reference, (Accessed 4 Feb 2022).

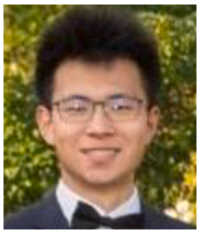
[59] Department for Levelling Up, Housing & Communities, Energy performance of buildings data, 2022, URL https://epc.opendatacommunities.org/docs/api, (Accessed 24 Feb 2022).

[60] HM Land Registry, HM land registry open data, 2022, URL https://landregistry.data.gov.uk/, (Accessed 13 Oct 2022).

[61] B. Glimm, I. Horrocks, B. Motik, G. Stoilos, Z. Wang, HermiT: An OWL 2 reasoner, J. Autom. Reason. 53 (3) (2014) 245–269, http://dx.doi.org/10.1007/s10817-014-9305-1.

[62] K. Schekotihin, P. Rodler, W. Schmid, OntoDebug: Interactive Ontology Debugging Plug-in for Protégé, in: Lecture Notes in Computer Science, Springer International Publishing, 2018, pp. 340–359, http://dx.doi.org/10.1007/978-3-319-90050-6_19.

[63] M.A. Musen, Protégé Team, The Protégé project: A look back and a look forward, AI Matters 1 (4) (2015) 4–12, http://dx.doi.org/10.1145/2757001.2757003.



Jiaru Bai received a B.Eng. in Chemical Engineering from the Dalian University of Technology and the University of Manchester and a M.Phil. in Advanced Chemical Engineering from the University of Cambridge. He is a Ph.D. student at the University of Cambridge working on automating chemical research using knowledge graphs.

Kok Foong Lee holds a Ph.D. in Chemical Engineering from the University of Cambridge and M.Eng in Chemical Engineering from the University of Birmingham. He is currently working on developing knowledge graph technologies as a computational scientist at CMCL Innovations.

Markus Hofmeister received a B.Sc. in Process Engineering from Technical University Freiberg (Germany) and a Dipl.-Ing. in Reservoir Engineering from Montanuniversitaet Leoben (Austria). He is currently pursuing a Ph.D. at the University of Cambridge working on semantic web technology applications, mainly in the energy and smart city domain.

Sebastian Mosbach obtained a Master's degree in Theoretical Physics and a Ph.D. in Chemical Engineering from the University of Cambridge. He is a Senior Research Associate at the University of Cambridge and a Principal Scientist at CMCL Innovations working on the development of digital twins.

Jethro Akroyd obtained his M.Eng. and Ph.D. in Chemical Engineering from the University of Cambridge. He is Senior Research Associate at the University of Cambridge and Principal Engineer at CMCL Innovations working on the cross-domain interoperability of digital twins.

Markus Kraft obtained the degree Diplom-Technomathematiker at the University of Kaiserslautern and his Doctor rerum naturalium in Technical Chemistry at the same University. He is a Professor of Chemical Engineering at the University of Cambridge, and the director of CARES Ltd, the Singapore Cambridge CREATE Research Centre, where he works primarily on kinetic modelling and knowledge engineering.