2021 European Conference on Computing in Construction
Online eConference
July 26-28, 2021

An Application of the PROV-O Vocabulary to Building Energy Performance Simulation

Cathal Hoare[1], Sergio Pinheiro[1,2], James O'Donnell[1],
[1]School of Mechanical & Materials Engineering and UCD Energy Institute,
University College Dublin, Ireland
[2]BAM Ireland, Kill, Co. Kildare, Ireland

ABSTRACT

Like many technological advances, Building Energy Performance Simulation (BEPS) tools face barriers to their adaptation. Stakeholders must be able to trust results - building owners and occupiers want to know how the results will benefit them, while companies require a return in investment through the use of such tools. Furthermore, engineers using such systems should be able to examine their use over time to tune their accuracy. Provenance systems seek to capture information that can be used to fulfill these goals. An application of the W3C PROV-O vocabulary to record typical BEPS processes is described. An interface to support engineers when recording provenance data is then introduced.

INTRODUCTION

The building sector accounts for a significant portion of the overall energy demand and greenhouse gas emissions of any country. Around 39% and 36% of the CO_2 emissions are associated with buildings in the United States and Europe respectively (Energy for Europe by European Commission 2018). In Europe, buildings are responsible for 40% of overall energy consumption. One viable solution is to analyze the energy performance of buildings and use the results to inform various decision-making processes. While the body of work around this approach is well developed, many obstacles remain. Though willing to embrace new technologies, companies and government bodies are concerned by the accuracy of results (ECTP 2019). An environmental and energy performance gap between predicted and operational buildings is a widespread industry issue (de Wilde 2014). This can occur for a number of reasons, including quality of data input to the simulation process, as well as flaws in the process itself (Ali et al. 2019). A wider societal mistrust continues to grow around simulations and other data intensive approaches such as deep learning. This creates a hesitancy to share data and adopting results from these types of processes (González et al. 2019).

Data provenance helps promote trust and enables mechanisms to help users of BEPS modelling outputs to have confidence in those results (Zafar et al. 2017). The work described here would enable various parts of a BEPS process to be attributed to individuals and organizations as well as enable resproducability and analysis to improve accuracy. These concepts have been identified as key components for producing trust amongst consumers of modelling results (Stoica et al. 2017); that work identified explainable results as a key goal in the field of AI, Information Retrieval, Search and other fields. Consumers of results (technical experts through to building owners) must be confident that answers provided by decision support systems provide causal inferences rather than coincidental results. This is achieved by:

- being able to explain the source and quality of inputs to an Artificial Intelligence (AI) system;
- being able to identify the the properties of input that caused a result to be derived;
- being able to examine the effect of small perturbations to these inputs and allow engineers to experiment with these changes to assist in developing their understanding of the inputs, models and their outputs.

Stoica et al. (2017) further assert that a key requirement for supporting explainable simulations is ensuring their reproducability. To provide suitable metadata and mechanisms that admit these goals is to provide an explanation of a model's data and process provenance.

Data provenance, also known as data lineage or data pedigree is concerned with providing a history of the steps - including any data and process - used to derive some new data artifact. Provenance research has evolved in science and engineering, especially in areas under scrutiny, such as climate change, business where financial records are subject to formal review, engineering fields, such as aeronautical engineering, and food and pharmaceutical production (Glavic & Dittrich 2007). A search of research literature found few concrete references to data provenance in the construction domain; related fields, such as change management and version control, would, in the authors' opinion, inform a broader provenance view of BEPS. The overarching aim of this paper is to address this deficit by exploring how the PROV-O linked data vocabulary can be applied to record BEPS provenance.

The paper continues by reviewing existing work in the area of provenance, and in particular, its application to BEPS. Then, a typical BEPS process is described and the PROV-O provenance vocabulary is reviewed. A description of the application of PROV-O to record a BEPS process is then provided. The paper will continue by describing an interface that was developed to simplify recording BEPS provenance.

BACKGROUND

A review of the literature for the Architecture, Engineering and Construction (AEC) domain finds few references to work on provenance. However, provenance has been applied in science and engineering domains such as scientific simulation, food production and business (Simmhan et al. 2005a,b). This section seeks to place these works in the context of provenance, discussing their relevance alongside references to related work on provenance in other domains. Data and process provenance has several purposes. Within the AEC domain, these include Audit/Attribution, Replication and Data Quality. Audit/Attribution is the ability to learn about the source of information, including, but not limited to the ability to know who created information and both how and when this occurred (Ambrosio et al. 2017). This application is not just about attributing information to individuals, but, for example, can also be used to trace faults with measuring equipment. Replication is the ability to recreate a process (Zafar et al. 2017), and is used to examine the execution of a simulation, perhaps to compare with real world data in order to refine the accuracy of the process for future applications (Leschert & Mclean 2015).

Several approaches have been explored to provide provenance. These range from extensions to databases to the use of linked data based approaches. While some focused on describing provenance of data or process (Hasham et al. 2018), we focus on an encompassing solution in order to address all three purposes discussed earlier. Several provenance standards have been proposed. Because of their broad applicability, and ability to admit reasoning, we focus on linked data based provenance approaches as described by (Moreau et al. 2015). A range of other implementations approaches have been taken to support provenance, including the use of Blockchain (Liang et al. 2017) and bespoke metadata schema for high performance computing environments (Dai et al. 2019). Indeed, scalability and operational overheads has received extensive attention, for example, Suh et al. (2019); while these approaches have merit, we believe that linked data provides a solution that is easier to implement and maintain in the context of the BEPS application context.

While many works focus on the process of creating and managing provenance data, the ability to reason about this information is key to answering more complex questions. Semantic and linked data approaches are particularly suitable for this purpose. Made up of structured data, annotated using the Resource Description Framework (RDF) standard and ontologies expressed using Web Ontology Language (OWL), that provides meaning for the RDF annotations, these data can be reasoned about both in terms of its values and the relationships between these. The PROV-O ontology (PROV-O: The PROV Ontology 2013), described in greater detail in the next sections, is provided by W3C as a baseline standard for expressing provenance information. Several other projects have explored the provision of provenance data using linked data approaches. Ornelas et al. describe the development and use of Open Provenance Model Ontology-e (OPMO-e) ontology for describing and reasoning about large scale scientific experiments (Ornelas et al. (2018)). Shaon et al. describe how the Digital Object Identifier (DOI) ontology was combined with others such as the data re-use standard, OAI-ORE, to provide the ability to reason about research that was input to a report in order to support the veracity of climate research (Shaon et al. 2012). Leschert et al. describe the development of a linked-data approach to capture and reason about the provenance of power systems (Leschert & Mclean 2015); this is used to audit these systems to maintain, tune and enhance their operation.

SUMMARY OF ENERGY MODELLING PROCESS

The energy modelling process is composed of entities and relationships that describe an energy modelling session. This was derived both from interviews with energy modelling experts and with reference to IEA EBC Annex 60 project which promoted research and development of new methods to help design and operate energy efficient buildings and communities (Wetter & Van Treek 2017). The steps consist of:

- *Step 1 - Data Collection:* data collection is perhaps the most variable step in data modelling. Information can originate from a variety of sources, including drawings, survey information and sensor data. The collected data, the process and tools used to collect these data along with the agent that conducted the measurements are all variables that influence the accuracy of this step, and so should be recorded. Furthermore, data collection is implicitly part of the energy modelling process, but may be carried out independently - both in terms of process and time - from the actual modelling process;

- *Step 2 - BIM-based CAD Tool:* this tool is used to collate collected information. Driven by a software process, details of the software and hardware platform should be recorded to ensure that the model is reproducible. Any changes made to the data in the BIM should also be recorded;

- *Step 3 - Export Tool:* this is also a software artifact. It is used to export a defined subset of data from step 2. Both the software version and the configuration of the platform it runs on, along with any run-time settings can effect the model's outputs. While largely a software step, the version of step 2's data that the process was applied to should be recorded to ensure traceability;
- *Step 4 - Data Transformations:* various transformations can be applied to the exported data, fundamentally changing its character, and so influencing the model's output. Recorded details would include for example, procedures, software versions and its configuration;
- *Step 5 - Quality Control:* various quality control steps can be applied to the exported data, fundamentally changing its character, and so influencing the model's output. Details recorded include procedures followed and evaluation criteria;
- *Step 6 - Import Tool:* this step is heavily influenced by software. Like the earlier export step, both the software version and the configuration of the platform it runs on, along with any run-time settings can impact the output of the overall modelling process;
- *Step 7 - Simulations Tool (e.g. EnergyPlus):* similar to steps 2, 3 and 6, software version, hardware details and settings should be recorded to ensure traceability;
- *Step 8 - Outputs:* the output of the modelling process may be used as input to other processes, or appear in reports or documentation. In order to allow these uses to be traced back to the modelling process, they should be properly attributed to the modelling instance.

This description shows that BEPS modelling is a complex collaborative workflow, with multi-facetted data, multiple human agents and complex software running on multiple hardware platforms.

The PROV-O VOCABULARY

W3C's PROVPROV-O: The PROV Ontology (2013) is a collection of 12 documents that describes standards that seek to fulfill the requirements of the Provenance Incubator Group. At its core is the PROV-O ontology (shown in Figure 1) that describes 3 classes, and 7 fundamental relationships between these; ancillary classes and relationships are also defined, such as use of datetime to track temporal aspects provenance. Within the ontology, *entities* describe physical, digital or other concepts. The ontology extends the term to include collections, bundles and plans that represent collections of entities, entities that describe provenance and entities that describe ordered sets of actions that allow *agents* to achieve some goal. Entities are further expanded through the use of intra-entity relationships. Agents describe someone or something that trigger an activity resulting in the creation of or changes to entities. The PROV-O standard provides several sub-classes of agent including Person, Organization and Software Agent. An activity is something that acts upon an entity and is triggered by an agent. These classes are quite flexible in structure, consisting of a class ID and an optional set of attributes; this structure provides great flexibility when recording provenance.

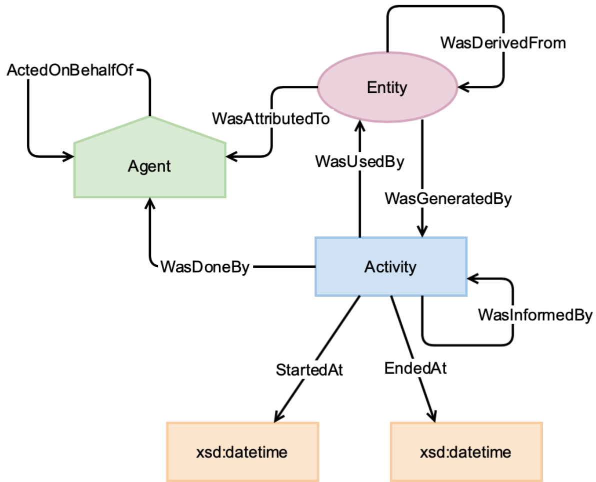

*Figure 1: W3C's Prov-O (*PROV-O: The PROV Ontology *(2013)) Ontology for Provenance Management.*

The classes are related through a set of relationships. Activities can use an entity or be informed by other activities. They also have a *begin* and *finish* time. Entities are derived from one another and can be generated by activities.

This ontology can be used to record and, subsequently, answer questions such as who carried out some activity, or what was the initial information used by some activity. When several activities are chained into a plan, the ontology allows for a rich understanding of the process, and admits attribution, recreation of the process and data quality traceability back to original sources of information.

APPLYING PROV-O TO THE MODELLING PROCESS

The paper continues by examining how the PROV-O ontology is applied to create a representation for the provenance of each step in the modelling process. This representation provides a set of relationships between activities, agents and entities that can be used to audit or reproduce the modelling process, attribute data and modelling steps and provide a basis for measuring data quality. The steps, shown in Figure 2, while derived from detailed interview with BEPSs specialists, are illustrative, and other processes can be accommodated.

Steps 1 & 2 - Collecting Data and Importing to BIM

The provenance process begins by creating a representation of the simulation to provide context for the provenance data and any published results. We can say that the modelling process explicitly starts once the information is available, and indeed, when provenance is employed to reproduce the modelling process or to attribute actions in the process to some activity or agent, then tracing as far as data input is sufficient; data can be attributed to its source, and having data sets in their original form allows a simulation process to be rerun. In order to assess data quality, it is ideal to have each data set's provenance, including how the data was acquired, details of hardware used, qualifications of operators of equipment, and any other factors that might affect data quality. However, this information may not be available - third parties responsible for its measurement may not record such details.

The import process is described by the Import to BIM activity and involves the input data and updated BIM state (each an represented as an entity) and three agents representing the modeler, the company for which he works and the BIM server to which information is up-loaded. The company is responsible for maintaining the BIM server, while the modeler is responsible for triggering the import process. The BIM server is considered to have done the actual import. The modified BIM state is attributed to the three agents and can be said to have been generated by the import activity which used the input entities.

The only actual data source that persist in the BEPS process are the input data sources. Thereafter, the comprehensive view of processes applied, and recorded in the provenance system, will allow recreation of intermediate data states. Therefore, the focus is on the state of software or any manual steps taken by agents.

Step 3 - Exporting from BIM

The provenance for the next step of the process, involves describing how the collected data from the BIM server is exported to produce a data set that can be cleaned and imported to, in this case, EnergyPlus to execute the actual modelling process. At this point, two entities, the BIM's data set and the exported data are represented. These are acted upon by the export from BIM activity, and acted upon by four agents, including the BIM server, the export tool, the modeler and their company. The exported dataset is derived from the BIM's data. The activity uses the BIM's data set to generate the exported data set. The activity is informed by the BIM server and export tool software agents, while the activity is done by the modeler that triggers the export process. The exported dataset is attributed to the export tool, modeler and company. Finally, the company is responsible for the software agent's maintenance. The steps taken by the modeler, and especially the configuration of the export tool is the most important information recorded in this step.

Steps 4 & 5 - Transformation and Quality Control

Steps 4 and 5 deal with transformation and cleaning of the exported data. The provenance models for both steps are broadly similar consisting of two entities, the data input to the step, and the resultant output for the step. Each step has an associated step and is associated with three agents - the modeler carrying out the steps, their company (with its aforementioned responsibilities for maintaining software and platforms) and any tools used in the transformation and cleaning processes. As in earlier steps, the output data set is derived from the input set, while the activity uses the input to generate the output. The activity is informed by any software tools, while the activity itself is done by the modeler who triggers the use of software agents in pursuit of the goals of the task. The output is attributed to the three agents.

These steps differ from previous steps somewhat in that the activities are bespoke for the simulation being undertaken. Some of these may involve applying calculations to the data, substituting values, replacing values such as those from archetypes. Where these actions are standard, the actions taken along with references to company standards and procedures documents that describe the steps, should be recorded for the activity (process document). Where the actions are bespoke, each should be rigorously documented; the level of detail provided should be sufficient to allow another agent to reproduce the actions.

Step 6 - Importing to EnergyPlus

This step is largely software based. Again, there are two entities, the cleaned data, as input, and the EnergyPlus state after import. The activity is supported by four agents - two software, the import tool and EnergyPlus software, the modeler and the modeler's company. The import tool and modeler do the activity, and the activity is informed by the EnergyPlus software. EnergyPlus' state is informed (updated) by the import tool agent. Similar in structure to Step 3, the same relationships exist between the entities, the entities and activities, and the output is attributed to EnergyPlus, modeler and company. The steps taken by the modeler, and especially the configuration of the import tool is the most important information recorded in this step.

Step 7 - Simulation

The model for Step 7 is - like Steps 3 and 6 - largely software oriented. While the modeler agent is responsible for configuring the simulation engine and editing

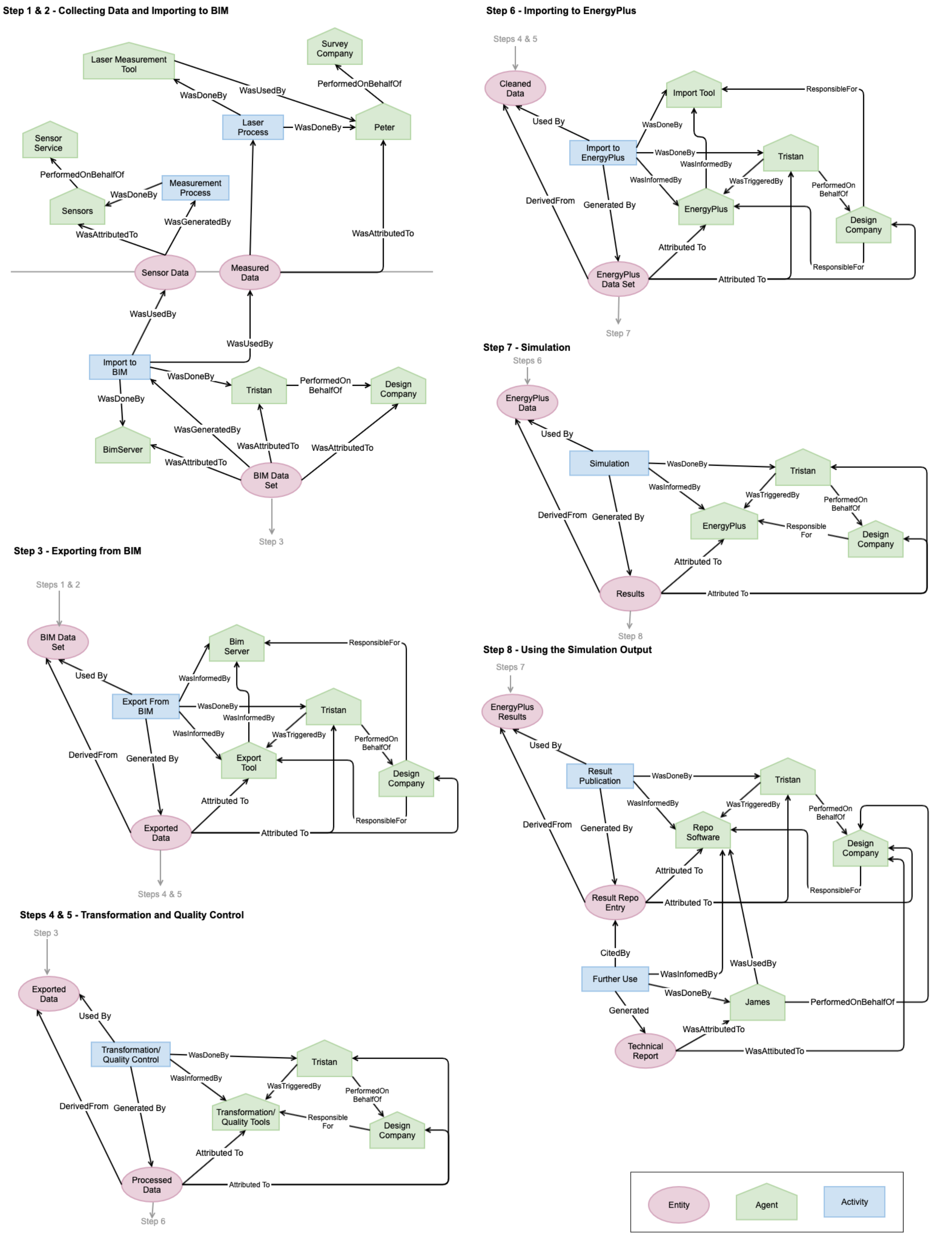

Figure 2: Prov-O Representation for all 8 BEPS Steps.

the editing the input file for simulation, it is the software itself that is best positioned to record its own state. Again, there are two entities - output derived from input, and associated with the simulation engine using the same relationships as before. Three agents, the EnergyPlus software, modeler and his/her company are also included, again sharing the same relationships as in previous steps. The steps taken by the modeler, and especially the configuration of the EnergyPlus tool is the most important information recorded in this step.

Step 8 - Using the Simulation Output

Step 8 describes how a result persists. The model described here assumes that some system assists with the tracking of results. Where this is the case, a representation of the result - the output file, or a document describing the results - would be published to the repository; the simulation object in the provenance model would be updated to reflect this. This final published entity is derived from the EnergyPlus result entity and generated by the publication entity. The action is triggered by the modeler on behalf of the company.

Step 8 is divided into explicit and implicit uses; this acts as a boundary between actions that are part of the BEPS modelling process and those that represent subsequent use of information generated by the process. This is captured by a generic Further Use activity. This activity generates some output by synthesizing many sources of information; therefore, this new entity is not derived from the modelling result. In this case the agent involved carrying out this task should cite the source - the modelling process' result - and update the result's citation list. This step creates a lineage from the entities that use the result through to the inputs to the modelling process.

IMPLEMENTING THE MODEL

We now present a prototype system for tracking data provenance during a BEPS process. The system is implemented as a standalone tool as it is recording a process that is implemented as a tool-chain that consumes multiple independent information sources. The user interface of the prototype is intentionally simplified - for example, allowing users to upload documents to describe activities or refer to online information sources - rather then explicitly asking for this information. This is for two reasons: first, it allows reuse of documentation, and second, it allows the prototype to describe a wide range of scenarios. The section will conclude by presenting how the prototype records information about each step of the BEPS modelling process.

An engineer would initially be presented with a pair of screens that organize tasks in a project component hierarchy. Each task is represented by information about the task and a collection of steps, as described earlier (see Figure 3). General data about the task includes its title, responsible person in the company, last modified data (and other metadata), and an overall quality score for the task. Each step that has been completed to date is listed in a table below. Each step is described by its metadata including a title, quality score, modified time and responsible person. Clicking on the row expands it to reveal a list of PROV-O relationships that describe the overall provenance for the step. Steps are related through their input and output data artifacts; for instance, Step 1's output is an input for step two. Relationships between entities within the step are described, for example the fact that the export tool was executed by a particular person is captured. Note the entity type - process, agent, source or output - is captured in abbreviated form in the triple. These triples are stored in an Apache Jena triple store and so can be subsequently reasoned about using SPARQL queries. This tabular view will be augmented at a later point with a graphical representation. New steps can be added by clicking on the appropriate button to add a new step. The first involves providing general metadata about the step, including title and responsible person. Any new processes, agents or data entities not previously represented on the system must then be added. The process description must always be added here as it will contain nuanced details that were not previously represented on the system. Other entities will usually have to be added - for example, the output of the step will generally be unique to the step and task. These details are added via pop up dialog boxes that capture the information described in Figures 4.

Adding details of a new step continues by prompting the user to describe the process by capturing relationships as triples. This is simplified by listing data entities, processes and agents associated with the task in drop-downs, along with provenance relationships, allowing users to quickly enter the required details.

CONCLUSIONS

BEPS holds demonstrable promise in contributing to the development of an energy efficient buildings and lowering energy emissions. These approaches suffer from several drawbacks in their implementation. Overall, there is a lack of confidence in such approaches. A performance gap often occurs between predicted energy consumption and that observed once a building is operational; this gap can occur for a number of reasons, including inaccurate inputs and missteps in the simulation process. The performance gap, concerns about value for money, along with other trust related issues reduces the adoption of simulation results by stakeholders, including the wider construction industry, decision makers and occupants.

This work posits that providing a system that provides auditable, replicable and explainable results would increase trust in BEPS modelling. By making results auditable, issues around quality can be resolved by investigating what data or process step caused gaps between predicted and actual outcomes; furthermore, the system allows for correction of other outputs where erroneous data sources or pro-

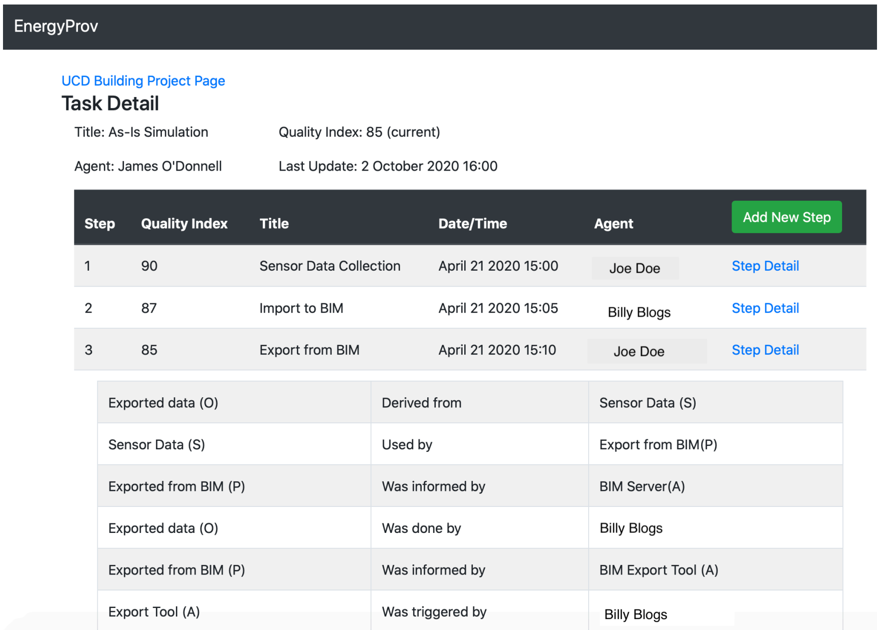

Figure 3: Interface that describes a simulation step, including a summary steps taken to complete the task. Each step can be viewed in detail by clicking on a row.

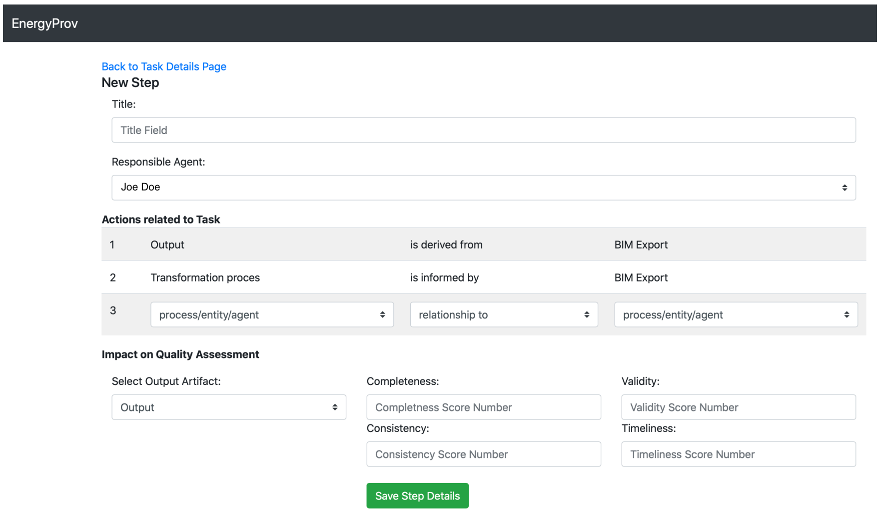

Figure 4: Interface used by an engineer to describe step in a simulation task, including an evaluation of the impact of this step on quality of the simulation.

cess outputs were used. Results are made explainable through the provision of a nuanced view of the process and data employed. Overtime, the system would provide feedback to improve design processes by identifying where errors occurred.

FUTURE WORK

While this work has significantly advanced the topic of provenance for BEPS, several outstanding issues remain. The human factors impact of using this type of tool must be investigated. Questions such as whether the approach provide burdensome for engineers, whether engineers will take the time to accurately record information, and how the quality of provenance information be measured and maintained remain open, and of significant importance. Approaches to assist in the analysis of provenance also remain open. While SPARQL provides an expressive means of accessing provenance information, its use is relatively specialized. Improved analysis and reporting tools are required to make the information captured by systems such as the one described here more accessible.

ACKNOWLEDGEMENTS

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under the SFI Strategic Partnership Programme Grant number SFI/15/SPP/E3125.

REFERENCES

Ali, U., Shamsi, M. H., Hoare, C., Mangina, E. & O'Donnell, J. (2019), 'A data-driven approach for multi-scale building archetypes development', Energy and Buildings **202**, 109364.

Ambrosio, L. M., David, J. M. N., Braga, R., Ströele, V., Campos, F. & Araújo, M. A. (2017), Wip: Prov-se-o: A provenance ontology to support scientists in scientific experimentation process, in '2017 IEEE/ACM 12th International Workshop on Software Engineering for Science (SE4Science)', pp. 15–21.

Dai, D., Chen, Y., Carns, P., Jenkins, J., Zhang, W. & Ross, R. (2019), 'Managing rich metadata in high-performance computing systems using a graph model', IEEE Transactions on Parallel and Distributed Systems **30**(7), 1613–1627.

de Wilde, P. (2014), 'The gap between predicted and measured energy performance of buildings: A framework for investigation', Automation in Construction **41**, 40 – 49.

ECTP (2019), 'Artificial intelligence – research and innovation needs – manufacturing, energy intensive industries, bio-based industries and construction.'.

Energy for Europe by European Commission (2018). Online; accessed 10-July-2021.
URL: https://ec.europa.eu/energy/en

Glavic, B. & Dittrich, K. (2007), Data provenance: A categorization of existing approaches., in 'Datenbanksysteme in Business, Technologie und Web', pp. 227–241.

González, F., Yu, Y., Figueroa, A., López, C. & Aragon, C. (2019), Global reactions to the cambridge analytica scandal: A cross-language social media study, in 'Companion Proceedings of The 2019 World Wide Web Conference', WWW '19, Association for Computing Machinery, New York, NY, USA, p. 799–806.

Hasham, K., Munir, K. & McClatchey, R. (2018), 'Cloud infrastructure provenance collection and management to reproduce scientific workflows execution', Future Generation Computer Systems **86**, 799 – 820.

Leschert, D. & Mclean, P. (2015), Power system model data provenance, in '2015 IEEE/IAS 51st Industrial Commercial Power Systems Technical Conference (I CPS)', pp. 1–9.

Liang, X., Shetty, S., Tosh, D., Kamhoua, C., Kwiat, K. & Njilla, L. (2017), Provchain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability, in 'Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing', CCGrid '17, IEEE Press, p. 468–477.

Moreau, L., Groth, P., Cheney, J., Lebo, T. & Miles, S. (2015), 'The rationale of prov', Journal of Web Semantics **35**, 235 – 257.

Ornelas, T., Braga, R., David, J. M. N., Campos, F. & Castro, G. (2018), 'Provenance data discovery through semantic web resources', Concurrency and Computation: Practice and Experience **30**(6), e4366. e4366 CPE-16-0291.R3.

PROV-O: The PROV Ontology (2013).
URL: https://www.w3.org/TR/prov-o/

Shaon, A., Callaghan, S., Lawrence, B., Matthews, B., Osborn, T., Harpham, C. & Woolf, A. (2012), 'Opening up climate research: A linked data approach to publishing data provenance', Int. J. Digit. Curation **7**, 163–173.

Simmhan, Y., Plale, B. & Gannon, D. (2005a), 'A survey of data provenance in e-science', SIGMOD Record **34**, 31–36.

Simmhan, Y., Plale, B. & Gannon, D. (2005b), 'A survey of data provenance in e-science', SIGMOD Record **34**, 31–36.

Stoica, I., Song, D., Popa, R. A., Patterson, D. A., Mahoney, M. W., Katz, R. H., Joseph, A. D., Jordan, M., Hellerstein, J. M., Gonzalez, J., Goldberg, K., Ghodsi, A., Culler, D. E. & Abbeel, P. (2017), A berkeley view of systems challenges for ai, Technical Report UCB/EECS-2017-159, EECS Department, University of California, Berkeley.
URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EE
2017-159.html

Suh, Y.-K., Lee, K. Y. & Baek, N. (2019), Pegasef: A provenance-based big data service framework for efficient simulation execution on shared computing clusters, in W. Lee & C. K. Leung, eds, 'Big Data Applications and Services 2017', Springer Singapore, Singapore, pp. 175–182.

Wetter, M. & Van Treek, C. (2017), IEA EBC Annex 60 Final Report, International Energy Agency (IEA), EBC Bookshop C/o AECOM Ltd The Colmore Building Colmore Circus Queensway Birmingham B4 6AT United Kingdom.
URL: http://www.iea-annex60.org/

Zafar, F., Khan, A., Suhail, S., Ahmed, I., Hameed, K., Khan, H. M., Jabeen, F. & Anjum, A. (2017), 'Trustworthy data: A survey, taxonomy and future trends of secure provenance schemes', Journal of Network and Computer Applications **94**, 50 – 68.