Reliability of Design Data Through Provenance Management

Tim G. Giese(✉) and Reiner Anderl

Technische Universität Darmstadt, 64287 Darmstadt, Germany
{giese,anderl}@dik.tu-darmstadt.de

Abstract. In today's virtual product development, huge amounts of data are shared and exchanged between a large number of experts in collaborative and highly complex design tasks. While in these processes designers are oftentimes working with data, which were not created by themselves, they are missing knowledge and transparency about the origin and reliability of the data source. Approaching this problem, we identified the necessity of a responsible and transparent generation and usage of design data. Therefore, we developed a concept, which tracks and stores the historical record of a data item and its modifications in order to identify and evaluate the source of the data item. The concept proposes a novel provenance model, which consists of a provenance graph, design criteria and evaluation criteria. To validate the concept, a prototypical implementation was conducted and evaluated. We came to the conclusion, that the presented concept can be used effectively to model and evaluate the historical record of a data set in the virtual product development in order to create a transparent and reliable use and generation of design data.

Keywords: Data provenance · Data literacy · Virtual product development

1 Introduction

In the modern product lifecycle management (PLM), data is seen as one of the most valuable assets. Massive amounts of data are collected, stored and distributed since the majority of the workforce in PLM are interacting with data on a daily level [1, 2]. However, research indicates that a vast number of workers are lacking essential skills to properly interact with data and use its full potential. This aspect is addressed by the research area Data Literacy [2]. Moreover, with the exchange of bigger amounts of data, processes in the design of virtual products are becoming more complex and untransparent. In global and collaborative projects, designers are often working with data, which were not created by them. Furthermore, there oftentimes exists missing knowledge about the origin and reliability of data they are working with [3]. This lack of transparency leads to a missing trust in the data and hence the potential of data is not fully used. Consequently, research proposes a transformation of the PLM towards a responsible generation and usage of design data to create a reliable and transparent

Published by Springer Nature Switzerland AG 2023
F. Noël et al. (Eds.): PLM 2022, IFIP AICT 667, pp. 274–283, 2023.
https://doi.org/10.1007/978-3-031-25182-5_27

product lifecycle. In order to do so, all participants of the product lifecycle have the responsibility to deliver reliable and high quality data [4]. However, suitable solutions in order to introduce Data Literacy to product development and provide designers with more competence in terms of gaining knowledge about a data source, are still under development [2].

To approach this issue and help designers not having to blindly rely on the quality of data, we developed a concept which provides a designer with reliable data in terms of identifying and evaluating the source of a data item. This concept consists of the development of a completely novel provenance model for virtual product development and focuses on tracking and storing the historical record of the generation and modification of data sets.

2 State of the Art

In the following, current research on Data Literacy and Data Provenance is described as well as the relevant previous work conducted by us.

2.1 Data Literacy

Although the research topic "Data Literacy" is tremendously gaining in significance, there does not exist a generally accepted definition of the term. However, what definitions have in common is the fact that the term is used to describe the ability to collect, critically assess and consciously apply data in a given context [5]. Nevertheless, this competence is not focused on particular scientific fields only, but being data literate rather describes an interdisciplinary data expertise [6]. The key competencies, which a data literate individual is considered to be equipped with are: exploration, prediction and inference. Using these competencies useful conclusions from large and diverse data sets can be drawn [6]. Despite the importance of these competencies a lack of these skills can be identified in the engineering workforce. Recent studies show that companies are increasingly looking for employees with an expertise in the interaction with data [7]. Furthermore, the skillset of a data literate individual is not limited to academia or the workforce only, but additionally Data Literacy focuses on skills in order to solve "real-world" problems as well. This is emphasized by the fact, that the varying definitions of Data Literacy additionally describe a skillset regarding the interaction of data, which can be used for everyday thinking and reasoning [5]. Since daily interactions with data are nowadays inevitable and common place throughout all age groups and all areas of life, Data Literacy is tremendously gaining in significance. Due to the ongoing digitalization, researchers highlight the necessity of knowledge about a proper interaction with data and even compare the importance of being data literate with the ability of how to read [5].

2.2 Data Provenance

Due to the increasing digitalization, the amount of data which is exchanged and shared between participants of the supply chain is significantly increasing. Furthermore, the

growing amount of participants often cause unclear situations about the origin or owner of a data item and its reliability [8]. The scientific area which is focused on the detailed description about the origin and authenticity of data is called "Data Provenance" or "Data Lineage". The provenance of a data item can be seen as the historical record of its derivation [9]. This record contains information about the source, processes and the current representation of the data item. Especially in collaborative, multidisciplinary projects (like the product lifecycle), in which designers often work with data which were not created by themselves, knowledge about the origin, transformations, which were applied to a certain data item and it's connection to other data sets are of great interest [8]. With knowledge about changed parameters, conducted simulations and individuals involved throughout the design process, the quality and reliability of a data item can be determined [3]. In literature models, which are able to save, depict and manage all decisions, procedures and results, which lead to the current state of a data set, are called provenance models [8].

There exist a great variety of different provenance models, which are able to receive the historical record of a data item. The majority of provenance models are graph-based since this creates the possibility to link data sets with defined relationships. Commonly well-known graph-based provenance models are the PROV W3C recommendation [10] and the Open Provenance Model [11]. The commonality of these models is, that they describe a meta provenance graph to describe the transformations a data item undergoes. These graphs generally consist of two categories: relationships (also called edges) and nodes (also called vertices) [10]. Relationships define the connection between two nodes. The second category – nodes – are further divided into 3 subcategories – entities, agents and activities [10]. Entities represent a physical, digital or conceptual unchanging state of a unit. Examples are documents, graphics or data sets. Activities describe an action which has been performed on or caused by an entity and are creating new entities while using already existing ones. An example is an action or a process. An agent can be understood to be taking on a role in an activity, which means, that an agent is responsible for an activity. Examples for an activity are a person, organization or software [10].

In comparison to provenance graphs, modern product data management systems (PDMS) are able to track and display the historical record of a data item in great detail as well. However, for tracking provenance information PDMS require that every contributor works on the same PDMS [12]. Once a data set leaves that PDMS the historical record of the data item cannot be tracked anymore [13]. Consequently, provenance models are needed next to PDMS since in the modern supply chain data is exchanged across company borders involving a great variety of different systems [12].

2.3 Previous Work

In the modern PLM an increasing amount of data is exchanged and shared among an increasing amount of participants [1]. This is oftentimes causing ambiguity of where a data item originates from and whether that source is reliable [3]. Even though in virtual product development the majority of the workforce are interacting with data on a daily basis, research indicates that, designers are missing skills and essential knowledge about the proper interaction with data originating from an external source in order to use that data to its fullest potential [3, 4, 7].

To approach this issue in a previous research paper [14], the authors developed a concept for the introduction and application of Data Literacy to virtual product development in order to create transparency and awareness for a responsible generation and use of design data [14]. This concept is called "Design Data Literacy" and the architecture is depicted in Fig. 1.

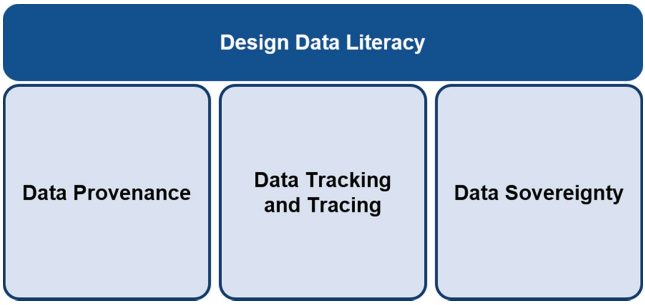

Fig. 1. Architecture of design data literacy [14]

In order to develop a comprehensive approach for the introduction of Data Literacy to virtual product development the concept is based on 3 major components – *Data Provenance*, *Data Tracking and Tracing*, as well as *Data Sovereignty*. Each component represents a different aspect of transparency and awareness and combined they convey the key competencies of Data Literacy adjusted to product development. The component Data Provenance is focused on identifying the origin and reliability of a data item. After a virtual product has been designed and further distributed, the component Data Tracking and Tracing is creating an overview about the receiver of the virtual product and the modifications which have been conducted. The third component – Data sovereignty – regulates the ownership and rights when the virtual product is further distributed [14].

In order to display and store relevant provenance information, a tree structure was identified to be best suited [14]. Since the development of the overall concept of Design Data Literacy, the component Data Provenance has been worked out in detail in terms of a development of a comprehensive data provenance model, which is described in the following.

3 Conceptual Approach

In order for designers in virtual product development to use design data to their fullest potential, they need to be provided knowledge about the origin and the quality of a data item. As identified by the authors in a previous research paper, more information about the historical record of a data set, which is exchanged across companies and several systems, can be achieved by a provenance model (see Fig. 1). The more knowledge a designer of a virtual product has about the origin and modifications of a data item, the more responsible and target-driven data can be used. Information about the origin and conditions under which a data set was produced generates high-quality and reliable design data without having to blindly rely on the creator.

Therefore, the goal of this concept is to create a comprehensive provenance model which generates a transparent and reliable use of design data within the product lifecycle. In order to achieve this goal three requirements are defined.

- R1: Provision of reliable and high-quality data
- R2: Identification and evaluation of the source of data
- R3: Transparent and responsible generation and usage of design data

To develop a comprehensive provenance model which addresses all of the requirements, the concept consists of 3 parts: A provenance graph, design criteria and evaluation criteria. The 3 parts are described in detail in the following.

3.1 Provenance Graph

To track the historical record of the derivation of a data item graph-based provenance models are well suited [3]. However, these models can be seen as meta models, which suggest a structure and are not yet adjusted to a certain scientific area or problem. In order to develop a provenance graph for virtual product development in which large numbers of experts from a great variety of disciplines are working collaboratively together, our model is mainly based on the PROV W3C recommendation [10], but has been extended by several additional structures. First, all relevant criteria to model and track the provenance of a data item in virtual product development were identified and subsequently integrated into a graph structure by linking the nodes with relationships (Fig. 2).

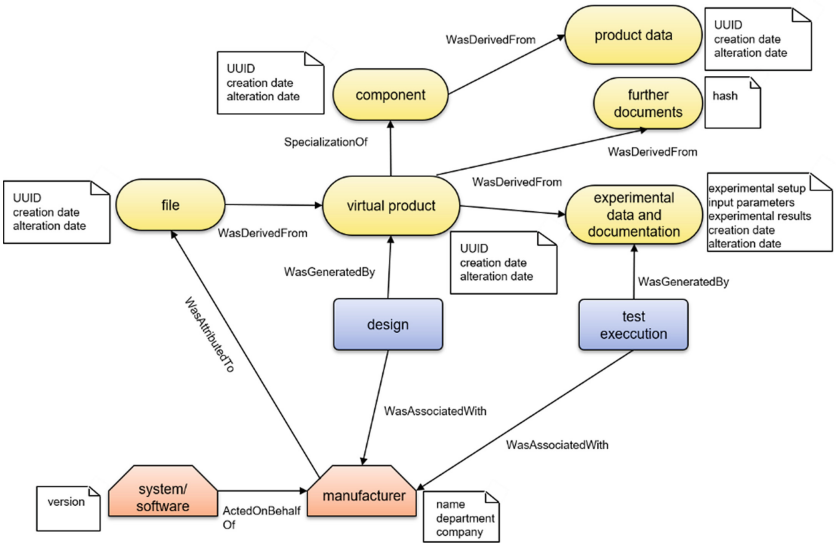

Fig. 2. Meta graph of presented provenance model (Color figure online)

For the model depicted above, the general layout of the PROV W3C recommendation has been adopted, consisting of entities, agents, activities and the associated relationships. The graph illustrates all relevant agents (orange), activities (blue) and entities (yellow), which are necessary in order to track provenance information in virtual product development. Additionally, attributes of a node are displayed in the white boxes. An agent represents a manufacturer or system involved in the design process. Moreover, the entity nodes are divided into 3 categories (file level, component level and product data level). The virtual product represents a 3D-data model of a product. The component node represents a part or an assembly of a virtual product. Moreover, product data describe the characteristics of a data item (e.g. material, density, weight, elastic module, surface, volume). Furthermore, entities which are necessary for a comprehensive description of a historical record of a data item are experimental data and documentation (documentation of the conducted simulations and their results) as well as further documents like measuring reports for example. The relevant actions involved in the design process of a virtual product are represented by the design and test execution nodes. Additionally, the relationships describe how 2 nodes in the design process are linked to each other. At last, for a clear identifiability of a data item and its current state the entities have a universally unique identifier (UUID) as well as creation and alteration dates, which are assigned to the nodes in terms of attributes.

The presented graph represents the structure of all relevant design data and their relationships to each other which are necessary to evaluate the quality and reliability of design data. To determine the origin and creator of a specific data item as well as transformations it underwent across its life cycle this graph still needs to be implemented and filled with specific information.

3.2 Design Criteria

Although with the presented provenance graph it is possible to track the historical record of a part or assembly, it can only display information which are stored in the file of the respective component. This means, if a previous designer did not specify from which standard a screw was taken for example (no entry in the file history), it is not possible to track this information. Consequently, this leads to the conclusion that a concept for a comprehensive, transparent and responsible use of design data cannot be guaranteed by usage of a provenance graph only. This highlights the necessity of further criteria, which need to be adhered to in order to guarantee a responsible design process. An extract of additional design criteria is depicted in Table 1.

For a responsible use of design data, every participating designer has the responsibility to create and promote trust. This means that conducted tests or simulations have to be documented for a subsequent designer, including time stamps, input parameters and test results. This creates the ability for a subsequent designer to re-run a test and therefore create trust in the data. Furthermore, inserting a PMI (Product Manufacturing Information) in a data model (consisting of a time stamp and information about the used standard) guarantees the ability to double check if a model was designed according to the correct standard and if that standard was still valid at that time. Moreover, in the product lifecycle documents like measuring reports are oftentimes exchanged among several

Table 1. Extract of design criteria

Design criterion 1	Verifiably documented tests and simulations
Design criterion 2	Test documentation must be available for subsequent designer
Design criterion 3	Standard-compliant design must be verifiable
Design criterion 4	External suppliers have to document provenance data
Design criterion 5	Further documents have to be securely encrypted

system participants. Securely encrypting these documents with hash values creates the advantage, that a recipient can verify, if the document was modified without permission.

The presented design criteria are included in the provenance graph as attributes of the respective entities (see Fig. 2).

3.3 Evaluation Criteria

In order to develop a data provenance model which automatically determines and evaluates the reliability of a source of a data item, evaluation criteria need to be identified. Therefore, the term reliability was defined first: "Reliability is the traceability of the history (place, author and time), conditions of origin and context under which a datum was created or modified in the product development process, so that the creation of the datum can be traced and reconstructed, and thus the quality of the data can be assessed, creating trust in the data". Subsequently, evaluation criteria were defined based on which algorithms search the provenance graph for a specific component. Following that, the reliability of the origin can be evaluated based on the selected criterion.

3.4 Architectural Approach

In this paragraph, the overall architecture of the provenance model is presented which is derived from the previous presented partial concepts (depicted in Fig. 3).

For the development of a comprehensive provenance model for virtual product development, it is crucial to consider the various phases of data processing: collection, storage, usage and transfer [4]. For the model presented in this paper, the phase of transfer is out of scope and the phase of usage is further divided into access and analysis. In order for a designer to determine whether the received data is reliable, the source of the data needs to be identified and evaluated. To achieve this, in the phase of collection a designer receives a data model, which is usually a virtual 3D-model of a product. With the usage of an Application Programming Interface (API) all relevant provenance data is exported in a machine-readable format and stored in a data base (in the presented concept a graph-data base is used and more specifically the provenance graph (see Sect. 3.1)). Additionally, further documents (like measuring reports) are encrypted with a hash value and the respective value is stored in the data base as well (phase storage). For a defined access to the data base a graphical user interface (GUI) was developed (phase access). In the GUI, a user can select a predefined evaluation criterion which results in an algorithm searching the graph for a specific part. Once the respective data item is found and the reliability

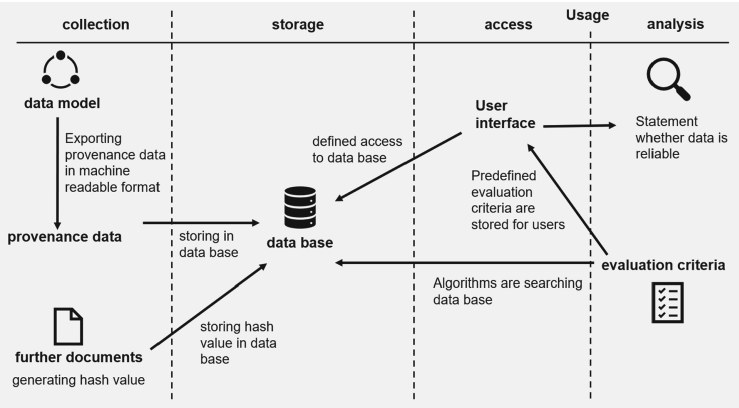

Fig. 3. Architecture of the overall provenance model

of its origin is evaluated information whether the source is reliable will be displayed in the GUI (phase analysis). As a specific application to the PLM domain, this provenance model can be used when a designer receives design data from a previous participant of the supply chain to determine where the data originates from, who was involved in creating the data and to rerun conducted experiments (like finite element, kinematic or dynamic analysis for example). The design data which the concept evaluates can be a file, virtual product or component of a product. Being able to determine the creator and conditions under which a data item was created provides designers with the possibility to evaluate the reliability and quality of data.

4 Validation

To validate the presented concept of a provenance model for virtual product development, the concept was prototypically implemented and the results are discussed.

4.1 Prototypical Implementation

For the implementation a sample virtual product was modelled in Siemens NX 12, consisting of parts modelled by several different designers and as a sample experiment a finite-element analysis was run on the product with predefined input parameters (the FEA is used as a sample experiment to retrieve its provenance information later on). After we received the 3D CAD model of the sample part, the provenance information had to be retrieved and exported from the data model. To achieve this, the API NX Open was used and the relevant provenance information exported into a CSV file. The application for the export was written in C#. Subsequently, the CSV file was imported into the graph data base using Neo4j (filling the provenance graph with specific information). Once the provenance information was stored in the data base an evaluation criterion could be selected from a drop-down menu in the developed GUI (which was written in Python 3). By selecting a criterion, an algorithm searched the graph data base for the respective

data item. For searching the graph, the search algorithms were written in Neo4j's graph query language Cypher. As an output the information from the provenance graph were displayed in the GUI (source (author) of the part, UUID, system and version the part was modelled in, time stamps of creation and alteration as well as input parameters and results of the FEA). At last, a message about the reliability of the origin was displayed based on the defined evaluation criterion.

4.2 Discussion

In the previous section it was demonstrated that the developed data provenance model can successfully identify and evaluate the origin of a data item (R1). With this knowledge the quality and reliability of a data item can be evaluated. Moreover, if the provenance information is passed on to a following designer or manufacturer reliable and high-quality design data is guaranteed (R2). Furthermore, keeping track of and displaying the historical record of every component used in a design creates a transparent and responsible generation and usage of design data (R3).

Even though we successfully developed a provenance model for the transparent and reliable use of design data within the product lifecycle, there are further aspects one has to take into account. Securely encrypting a document with a hash value only guarantees security to a certain degree. Even when assigned a proper digital signature, there always remains a risk that the hash value is forged and modified by a man in the middle [15]. Moreover, the goal of a transparent and reliable use of design data along the product lifecycle can only be guaranteed if all participants contribute to that [4]. When the design criteria proposed in Sect. 3.2 are adhered to this is ensured, however, the possibility remains that some individuals do not obey to these rules. In that case a transparent and reliable use of design data cannot be guaranteed anymore.

5 Conclusion

We successfully developed and implemented a novel concept for a transparent and reliable use and generation of design data along the product lifecycle by tracking and storing the historical record of any data item used in the design process of a virtual product. In the product lifecycle many experts from a great variety of disciplines are working collaboratively. However, the origin of a data item is often unclear and whether the data is reliable. The more complex and untransparent a design process gets, the more a designer of a virtual product has to blindly rely on the reliability of the data. However, if the quality of data is not known the potential cannot be fully used. To improve this issue, designers need to be provided with reliable and high-quality data. To approach this problem and provide this information our concept consists of a provenance graph, which tracks and stores the historical record of design data. With the provenance graph, it is possible to identify the origin and to understand which transformations a dataset underwent and who was responsible for the modifications. Moreover, to evaluate the quality and reliability of the data, evaluation criteria were developed. Using these criteria, the provenance graph can be searched for a specific data set and the reliability of the source and quality of the data is automatically evaluated. Additionally, to guarantee

a responsible generation and usage of design data along the whole product lifecycle, further design criteria were developed as a guideline of how to generate and exchange data with one another to achieve trust in data and a responsible and transparent product lifecycle.

Having considered all these aspects, we developed a comprehensive data provenance model for a responsible generation and use of design data in order to create transparency and reliability when exchanging data in virtual product development.

References

1. Collins, V., Lanz, J.: Managing data as an asset. CPA J. **89**, 22–27 (2019)
2. Pothier, W.G., Condon, P.B.: Towards data literacy competencies: business students, workforce needs, and the role of the librarian. J. Bus. Financ. Librariansh. **25**(3–4), 123–146 (2020). https://doi.org/10.1080/08963568.2019.1680189
3. Reitenbach, S., Vieweg, M., Hollmann, C., Becker, R.-G.: Usage of data provenance models in collaborative multi-disciplinary aero-engine design. In: Turbomachinery Technical Conference and Exposition (2020)
4. Die Bundesregierung – Bundeskanzleramt Deutschland: Datenstrategie der Bundesregierung (2021)
5. Frank, M., Walker, J., Attard, J., Tygel, A.: Data literacy - what is it and how can we make it happen? J. Commun. Inform. 4–8 (2016)
6. Adhikari, A., DeNero, J.: Computational and inferential thinking: the foundations of data science. University of California, Berkeley (2019)
7. Tech Partnership, Employer Insights: Skill Survey. https://www.techskills.org/globalassets/pdfs/research-2015/tec_employer_skill_survey_web.pdf
8. Glavic, B., Dittrich, K.R.: Data provenance: a categorization of existing approaches (2007)
9. Schreiber, A.: Provenance für workflows und prozesse, 4 March 2011
10. Groth, P., Moreau, L.: PROV-Overview: an overview of the PROV Family of documents. https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/. Accessed 17 Aug 2021
11. Moreau, L., Freire, J., Futurelle, J., McGrath, R.E., Myers, J., Paulson, P.: The open provenance model: an overview. In: Freire, J., Koop, D., Moreau, L. (eds.) Provenance and Annotation of Data and Processes. IPAW 2008. Lecture Notes in Computer Science, vol. 5272, pp. 323–326. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89965-5_31
12. Mesihovic, S., Malmqvist, J., Pikosz, P.: Product data management system-based support for engineering project management. J. Eng. Design **15**, 389–403 (2004)
13. Sebes, E.J., Stamp, M.: Solvable problems in enterprise digital rights management. Inf. Manag. Comput. Secur. **15**(1), 33–45 (2007). https://doi.org/10.1108/09685220710738769
14. Giese, T.G., Anderl, R.: Design data literacy – impact of data literacy in virtual product development. In: 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Brisbane, Australia, pp. 1–8 (2021)
15. Hasan, H.R., et al.: A blockchain-based approach for the creation of digital twins (2020)