Provenance Question-based AI Transparency and Accountable AI Governance

Laura Waltersdorfer[1,2] , **Dominique Hausler**[3] , **Tanja Auge**[3]
[1]Vienna University of Economics and Business, Austria
[2]University of Technology Vienna, Austria
[3]University of Regensburg, Germany
laura.waltersdorfer@wu.ac.at, {tanja.auge, dominique.hausler}@ur.de

Abstract

Ensuring transparency in *Artificial Intelligence (AI)* systems is critical for building trust and accountability. However, implementing technical governance and transparency in complex AI systems remains a challenge due to vague requirements, missing know-how and time resources. *Provenance questions (PQs)*, outlining transparency requirements of a system, can play a key role in counteracting this. Nevertheless, the implementation of technical transparency and suitable PQs in complex AI systems pose significant challenges. This paper presents an approach for the formalisation and transformation of PQs, aimed at improving AI system transparency. This involves a question analysis on a linguistic and provenance level, based on the W7 model. To this end, we propose two definitions for *simple* and *complex* PQs to map them to PROV-O concepts, followed by a discussion of a reference architecture.

1 Introduction

Artificial Intelligence (AI) is increasingly deployed in mission-critical domains, resulting in unintended consequences or even leading to negative event occurrences [Turri and Dzombak, 2023]. Responsible AI frameworks [UNESCO, 2021] and evolving regulation, for instance the European Union AI Act [European Commission, 2021], call for increased transparency mechanisms. System documentation and AI auditing are tools to counteract opaque system design, unclear data biases and fairness issues. However, AI system operators and engineers are often not aware of how to operationalise such governance frameworks. They usually do not know what information is critical for legal compliance or ethical concerns from a transparency perspective.

Provenance refers to tracing back the production processes of (digital) objects and is often considered as a technical means to establish transparency, accountability and trust. Despite the importance of provenance-enabled governance in AI, recent analysis of *Machine Learning (ML)* documentation remains vastly incomplete [Bhat *et al.*, 2023]. Provenance considerations are often not included in innovative AI systems described in research papers [Breit *et al.*, 2023], indicating a gap in technical implementation. In this context, *Provenance questions (PQ)* can be used to guide the design of transparent and responsible AI systems through structuring their transparency needs, e.g., as shown for information systems [Miles *et al.*, 2011]. Using PQs, e.g. *What was the evaluation procedure (...)?*, as transparency requirements can help to avoid ambiguities in what information is tracked or tracking unnecessary information leading to opacity, storage or data issues. Table 1 shows an overview of example PQs from different sources according to the W7 model, a workflow-specific provenance model [Ram and Liu, 2009]. Relevant PQs for AI systems can be derived from documentation templates for ML models such as Model Cards [Mitchell *et al.*, 2019] and Datasheets [Gebru *et al.*, 2018], white papers [Netherlands Court of Audit, 2021], other research prototypes, and data models [Fernandez *et al.*, 2023; Oppold and Herschel, 2020; Naja and et al., 2022], regulations [European Commission, 2021] or research [Liao *et al.*, 2020].

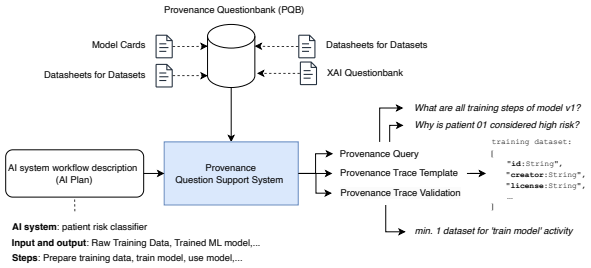

Figure 1: Approach of a provenance question framework for selecting PQs and generating provenance queries.

While these questions can guide the process of systematically collecting provenance requirements, there are multiple associated challenges:

- The selection of PQs is a time-consuming and high-effort, yet very critical process for defining provenance requirements.
- After collection, PQs need to be transformed into custom provenance data model and capturing mechanism.
- Domain experts are inexperienced in defining and transforming PQs into executable queries in query languages.

Example Provenance Question	Source
Who pre-processed the data in the dataset?	Data Sheets [Gebru *et al.*, 2018]
Who produced deployment guidelines (...)?	RAINS KG [Naja and et al., 2022]
Where is the source code stored?	Fides [Fernandez *et al.*, 2023]
When was the used dataset published?	Liquid Data Model [Oppold and Herschel, 2020]
How often does the system make errors?	XAI Questionbank [Liao *et al.*, 2020]
Why is this instance given this prediction?	XAI Questionbank [Liao *et al.*, 2020]
Which data and data sources does the algorithm use?	Whitepaper algorithmic audits [Netherlands Court of Audit, 2021]
What was the evaluation procedure (...)?	Model Cards [Mitchell *et al.*, 2019]
What is the intended use of the AI system?	EU AI Act [European Commission, 2021]
What is the license of the associated data?	RAINS KG [Naja and et al., 2022]

Table 1: Example provenance questions from different sources.

Contribution. To bridge this gap of provenance-related challenges in technical governance, we present a conceptual design for a provenance question support framework (cf. Figure 1). The framework takes as input an AI system workflow description and a provenance questionbank. The output are applicable PQs, trace templates, and trace validation criteria based on the AI system workflow description. Our aim is to enable different stakeholders during the AI system design to introduce provenance by design, i) to increase the overall transparency level of AI system by guided collection of audit traces, and ii) to increase the quality of provenance logs and traces.

Outline. We present an application scenario from explainable AI and give a brief overview on provenance data models and methods (cf. Section 2). We then define our PQs based on the W7-model (cf. Section 3), followed by question analysis on a linguistic and provenance level (cf. Section 4). We continue with a discussion on a reference architecture (cf. Section 5), related work (cf. Section 6), and conclude our approach (cf. Section 7).

2 Application Scenario & Background

Enabling technical transparency within a complex AI system is challenging, due to storage, efficiency or legal restrictions, preventing comprehensive collection of data. To achieve collecting relevant traces, PQs should be adapted to the AI system context. We consider the following scenario to illustrate the use of PQs.

2.1 Scenario Description

A hospital uses an AI system to predict health-adverse effects based on high-risk patient's medical data and demographic-social information [Liao *et al.*, 2021a]. Clinical practitioners use explanations to better understand the system outputs while providing healthcare services. Relevant PQs to increase the explainability of the the AI system might be:

PQ1 : What was the training process for the model?

PQ2 : Why is this patient considered to be high risk?

These natural language PQs are partially more adapted to the medical system context than the ones in Table 1. However, they conceal a degree of complexity before being able to be answered. *PQ1* may address different aspects: i) used entities (training datasets, ML models, used hyperparameters, ...), ii) specific activities (data cleaning or augmentation), iii) involved agents (people, organisations or software tools). While *PQ1* is a typical provenance question that can be modeled with common provenance models, *PQ2* is focused on explainability aspects, still being an open challenge in AI research. There are multiple approaches to achieve such explanations, for example counterfactuals, local feature contribution or system rules [Liao *et al.*, 2021b]. *PQ2* aims to understand the decision-making process of the model on a local level, this means understanding contributing features and thresholds for the high-risk class. Capturing provenance information for these questions is possible, but depends on the AI system design and implementation as well as the choice of provenance data model and capturing mechanism.

Interim Conclusion. In order to be able to i) come up with such suitable PQs, ii) adapt these PQs to the AI system context, and iii) transform them into executable queries, we require three main components (cf. Figure 1):

1. Questionbank collecting relevant example PQs.
2. Provenance data model depicting the AI system workflow description.
3. Method to derive provenance trace templates from PQs including transformation and execution of provenance queries.

We focus in this work on provenance data models (2.) and how to derive trace templates from PQs (3.), after already having collected relevant example PQs based on [Liao *et al.*, 2020; Naja and et al., 2022] for the questionbank (1.).

2.2 Provenance Data Model

The *AI system workflow description* defines the algorithmic processors and dataflows, e.g., the number of applied AI models and training data; it is used to derive expected provenance traces. For representing the workflow description, we reuse a generic and well-structured data model for recording workflow provenance *P-Plan ontology* [Garijo and Gil, 2012], extending the *PROV-O W3C recommendation* [Lebo *et al.*, 2013].

PROV-O is used to represent provenance information by providing the concepts of entities (`prov:Entity`), activities (`prov:Activity`), agents (`prov:Agent`). All concepts can be linked to each other through specific relation-

ships, e.g., `prov:wasGeneratedBy` for linking entities to activities. Whereas PROV-O models the actual execution traces, P-Plan extends the data model through concepts to model the planned execution of a workflow. To achieve this, plans (`pplan:Plan`), steps (`pplan:Step`), variables (`pplan:Variable`) and their relationships, e.g., `pplan:correspondsToStep` linking steps to activities. In the following, we present a simple representation of the medical AI system from the application scenario above.

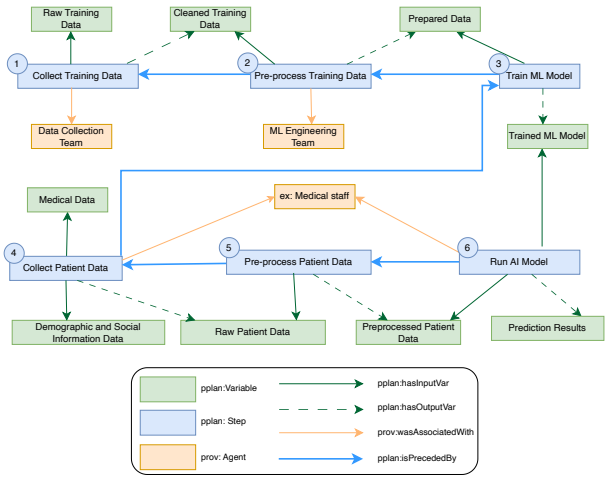

Figure 2: Example representation of AI system workflow description with planned activities `pplan:Steps` in blue, in- and output `pplan:Variables` in green and agents `prov:Agents` in orange using *P-Plan* ontology.

Figure 2 shows the AI system workflow description. The description consists of six planned activities (in blue), nine input and output data (in green) and three agents (in orange). For the system description of our application scenario, the three types of modelled concepts are `pplan:Step`, `pplan:Variable` and `prov:Agent`. In the following, we describe each planned activity, inputs and outputs:

1. *Collect Training Data*, having *Raw Training Data* as input and *Cleaned Training Data* as output, this step is related to *Data Collection Team* as (`prov:Agent`), with the property (`prov:wasAssociatedWith`).
2. *Preprocess Training Data*, having *Cleaned Training Data* as input and *Prepared Training Data* as output, this step is also associated with *ML Engineering Team*.
3. Followed by, *Train Model*, having *Prepared Training Data* now as input and outputting the *Trained Model*.
4. Then, when the medical staff is using the system, they first *Collect Patient Data*, gathering both *Demographic and Social Information*, as well as *Medical Data* from the patient and *Raw Data* as output combining them.
5. Afterwards, *Pre-process Patient Data* having *Raw Data* as input and *Preprocessed Data* as output.
6. Finally, the step *Run AI Model* [1], is executed by the medical staff using *Preprocessed data* as input and obtaining *Prediction Results*, among them being the risk class of the patient.

For each variable (`pplan:Variable`), relevant data provenance properties need to be collected. Example characteristics of the training data are identifiers, source, contents and license. Additionally, for each step (`pplan:Step`) workflow provenance is required, such as timestamp, the final status of the executed activity and agent IDs. To streamline the collection of provenance traces, we discuss provenance trace templates and transformation next.

Interim Conclusion. One key difficulty in complex AI systems is that required provenance data is spread across various stakeholders, in different data formats and involving different systems, making data identification, integration, mapping and storage challenging. This prevents a holistic overview required to answer provenance questions.

2.3 Provenance Trace Templates

In order to answer *PQ1*, different provenance data needs to be collected. This includes inputs, outputs and activities related to the training process. We assume that one input training data in the application scenario is the openly available MIMIC-III dataset [Johnson *et al.*, 2016]. The collected demographic and medical properties for each patient include age and blood pressure, represented in SNOMED CT, a clinical terminology system [Stearns *et al.*, 2001]. In our approach, we provide concrete *provenance trace templates* to structure the provenance collection process. It consists of key-value pairs {"<attribute-name>":"<datatype>"} of expected provenance characteristics. In step 1 *Collect Training*, the input (`prov:Entity`) corresponding to the `pplan:Variable` *Raw Training Data*, can be described in the following `JSON` trace file:

```
{
  "id": "MIMIC_III_Database_2016",
  "title": "MIMIC-III Clinical Database",
  "creator": "MIT Lab for Computational Physiology",
  "issued": "2016-01-01",
  "license": "https://physionet.org/about/licenses/"
}
```

The key consists of basic provenance data such as `id` and `title`. The value represents the provenance data. Templates are based on selected PQs and application domain. While some template components are consistent (e.g., timestamps), other components e.g., `license` are derived from additional PQs, such as *What is the license type of the used data?* The output of step 2 *Pre-process Training Data* is the cleaned and prepared data based on the original MIMIC III, and then input to *Train Model* step 3.

Even though this is not the full provenance information for *PQ1*, we can already answer selected PQs such as *When was the used dataset published?* (cf. Table 1). An approach to retrieve the training process information could be adapted to

[1] In the context of the scenario, we assume only one AI model is used, but multiple models could also be represented to reflect that different result properties stem from different models.

the chosen implementation technology to execute the following example provenance query:

```
SELECT Activity, Used, Generated, AttributedTo
FROM ProvenanceTable
WHERE Activity
    IN ('Preprocess_Data_Activity','Train_Model_Activity')
```

To answer *PQ2*, we need to analyse provenance traces backwards from the result of step (6) *Run AI Model*. After the model has classified a certain patient as high-risk, we want to find out the contributing features to this classification result. One way is to calculate *SHapley Additive exPlanations* (SHAP) values [Lundberg and Lee, 2017] after model training, the contribution of individual features to the models output. The following values indicate the importance of each feature for our scenario:

```
{
    "snomed:75367002": "0.64", #blood_pressure
    "snomed:424144002": "0.31", #age
    ...
}
```

The integration of provenance data from *PQ1* evaluates the training process. Other concepts such as local or counterfactual explainability methods [Wexler *et al.*, 2019] can be integrated to provide more context to the answer of *PQ2*.

Additional potential queries inquire involved agents, preprocessing steps or evaluation details. We require additional provenance data on pre-processing steps, additional used datasets (if applicable) or ML model and (training and evaluation) parameters, to answer PQs in more detail [Auge *et al.*, 2024].

Interim Conclusion. The advantage of providing provenance queries and trace templates is that relevant stakeholders know which information is needed (e.g., name, license, preprocessing steps and contributing factors). Already, through this simple example we observe the complexity of identifying the required provenance information from PQs to derive adequate trace templates across an AI system.

3 Provenance Question Definition

To formalise AI-specific PQs and to support provenance-by-design, we introduce the commonly-used W7 model [Ram and Liu, 2009]. We define the following general concepts: *simple* and *complex* PQs and provenance results.

W7 Model. In the W7 model, provenance is conceptualized as a combination of seven interconnected components to track events that affect data during its lifetime [Ram and Liu, 2009]. Different aspects of an event can be described by the following parameters: **`when`**, **`why`**, **`how`**, **`who`**, **`where`**, and **`which`** (cf. Figure 3). **`When`** refers to the time at which the event occurred. **`Why`** represents the reasons for the event. **`How`** describes the action leading up to the event. **`Who`** denotes the agents involved in the event. **`Where`** indicates the location of the event. **`Which`** specifies the programs or instruments used in the event. Furthermore, **`what`** denotes the event that affected the analysed data.

Provenance Question. The answer of a PQ provides information about the origin, processes, and context behind the generation of data or decisions in a system. We define a provenance question Q_P as tuple consisting of at least one question word defined in the W7 model, an object of interest and a (main) verb describing the activity. We distinguish two types of PQs, simple and (positive) complex:

Definition 1 (Simple provenance question)**.** *A **simple provenance question** Q_P is defined as a tuple (w, o, v), where $w \in \{$**`where`**, **`when`**, **`who`**$\}$ is a question word, o is the object of interest – a subject or a phrase – and v is a (main) verb describing the action.*

Definition 2 ((Positive) complex provenance question)**.** *A **(positive) complex provenance question** Q_P is a tuple (w, o_1, o_2, v), where $w \in \{$**`what`**, **`which`**, **`why`**, **`how`**$\}$ is a question word, o_1 is the object of interest – a subject or phrase –, o_2 is a second optional object and v is a (main) verb describing the action.*

Each simple or complex PQ can be extended by a condition c, which specifies the question in more detail such as time or location, e.g., *Who updated the training data last month?* or *How long does the AI process last?* (cf. Table 2). Negative PQs contain the extension **NOT**. However, these only apply to certain question words such as **`why`** or **`what`**, e.g., *Why is the instance not predicted to be a different outcome?* For other PQs such as *Who updated the training data?* or *When did the ML training process start?*, negation does not make sense in terms of content. In addition, not every **`what`**-question such as *What does the system output mean?* is a PQ. There are also PQs that do not begin with one of the seven question words, such as *Is [feature X] used or not used for the predictions?* A compound question (cf. *When did the AI process stop and start?*) cannot be answered directly, as two questions are asked in one. Consequently, it has to be split into two PQs.

Provenance Answers. A provenance answer is defined as cause, location, agent, time, instrument, reason, and more. Each question word has a corresponding class of answers. For example, a **`who`**-question is always answered by providing an agent, which in turn can be a living or non-living entity such as a system, an institution or a person (cf. Figure 3).

Definition 3 (Provenance result)**.** *Let o be an object of interest and Q_P be a simple or complex provenance question. A **provenance result** R_P is a string, number, file, list or figure specifying cause, location, agent, time, instrument, reason, and more. A provenance mapping is a function/homomorphism p mapping the provenance result R_P of o.*

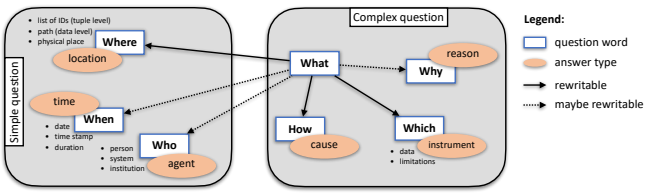

Figure 3: W7 model including provenance question word, answer type and possible answers.

		Example Question	Answer type
simple PQs	Q_1:	Who updated the training data in the last month ?	agent
	Q_2:	Who was involved in generating the data ?	agent, $\hookrightarrow Q'_2$
	Q_3:	When did the ML training process start ? When did it stop ?	time
	Q_4:	Where is the test data stored ?	location
complex PQs	Q_5:	Which input data sources have spatial/temporal resolution ?	instrument
	Q_6 :	How long does the ML training process take in the last run ?	time
	Q_7:	What are the limitations of the system ?	$\hookrightarrow Q'_7$ or Q''_7
	Q_8:	What is the source of the training data ?	$\hookrightarrow Q'_8$ or Q''_8
	Q_9:	What kind of output does the system give ?	—
	$Q_{10,1}$:	How often does the system make mistakes ?	cause
	$Q_{10,2}$:	How to improve the system ?	cause
	Q_{11}:	How does the system make predictions ?	cause
	$Q_{12:}$	Why is this patient considered to be high risk ?	reason
	Q_{13}:	Why is this instance not predicted to be a different outcome ?	reason
	Q_{14}:	What does the system output mean ?	no typical PQ
rewritten PQs	Q'_2:	Who generated the data ?	agent
	Q'_7	Which limitations does the system have ?	instrument
	Q''_7:	How is the system limited ?	cause
	Q'_8:	Where does the training data come from ?	location
	Q''_8:	What kind of data was the system trained on ?	—

Table 2: Example provenance questions of different structures including a question word, a (main) verb, and a subject and/or a phrase. If necessary, additional conditions and refinement can be added and PQs can be rewritten ($\hookrightarrow$).

Application Scenario. In the case of the (positive) complex *PQ2*, the provenance answer type is a *reason* explaining the classification result. Possible provenance results are i) the relevance of blood pressure (64%) and age parameter (31%) on the global model, ii) the relevance of blood pressure (64%) and age parameter (31%) for a concrete patient instance or iii) the concrete blood pressure and age values. The selection of the appropriate result depends on the background and intention of the system user asking the question. *PQ1* being a **`what`**-question makes it difficult to answer. Example provenance answer results can be a list of executed training activities and entities (e.g., input datasets) with relevant parameters, such as data splits or used algorithms.

4 Question Analysis

In order to tackle the challenge of ambiguous answer categories as described above, question analysis and rewriting is an essential part. This question transformation is needed to answer PQs based on the desired question intention and output information for our approach. We present selected example questions from explainable AI [Liao *et al.*, 2020] (cf. Table 2) to illustrate the linguistic rewriting and mapping to provenance concepts.

Question Selection. In Table 2, PQs are grouped into question type (simple, complex or rewritten), example questions and expected answer type. Questions Q_7-Q_{14} are derived from a question-driven design process for explainable AI [Liao *et al.*, 2020], covering aspects from output data to model performance, while rewritten queries are result of linguistic analysis based on the original ones.

Linguistic Rewriting to Provenance Questions. The rewriting consists of two levels: a linguistic and a provenance level. A syntactic analysis is performed to identify structural similarities. Each element of a sentence can consist of either one or more words and includes: subject (S), predicate (P), different kinds of phrases, objects and adverbials.

Linguistic Level. English follows the *subject-verb-object (SVO)* word order. Each sample question in Table 2 represents a *wh*-question, starting with a question word; **`when`**, **`why`**, **`how`**, **`who`**, **`where`**, **`which`** and **`how`** (cf. Figure 3). There are two different types of questions: *subject question* (cf. Q_2) and *object question* (cf. Q_4). The subject takes the initial position in a subject question and follows the SVO word order [Westergaard, 2009; Stromswold, 1995]. In object questions the auxiliary or a modal verb is postioned before the subject. At the linguistic level, the auxiliary and the main verb are defined as complex predicate.

Provenance Level. We differentiate between simple and complex PQs to provide adequate answers. Simple PQs

	Q_2			Q'_2		
	Who	was involved	in generating the data ?	Who	generated	the data ?
Linguistic Level	S	P	prepositional phrase	S	P	accusative object (NP_{Akk})
Provenance Level	question word	main verb	object of interest	question word	main verb	object of interest

Table 3: Example of provenance question rewriting.

start with **where**, **who**, or **when**. Those consist of a question word, a (main) verb and a phrase or subject.

Complex PQs start with **which**, **what**, **how**, or **why**. These include an object of interest (subject or phrase) and a second optional object such as a refinement. A refinement linguistically signals that the question word is part of a *wh*-object, narrowing and specifying what the question word asks for. As the answer to a **what**-question is always more general than to a **which**-question, we propose rewriting them to give more adequate answers. For Q_7, for example, the answer type can be defined as instrument or cause depending on the chosen rewritten form (cf. $Q_7 \hookrightarrow Q'_7, Q''_7$ in Table 2).

For provenance, only minimal working examples are considered. Rewriting is possible, as all important information is still part of the PQ and the answers will be of the same value. Most of the typical PQs are object questions, meaning the auxiliary or modal verb is placed right before the subject and the (main) verb is positioned at the end. Only the main verb is taken into account at this level. E.g., Q_2 is rewritten to Q'_2, following the same structure as Q_1. Table 3 illustrates the differences between the two levels discussed above.

PQs can be extended by conditions which further specify what is asked, representing a filter function (cf. Q_1, Q_6). Compound questions cannot be directly answered with provenance, they need to be rewritten (in two questions) like in Q_3. Some two-sentence PQs can build a causal chain e.g., $Q_{10,1}$ and $Q_{10,2}$, illustrating a special case of a compound question.

Mapping to Provenance Concepts. We map the question answers to ontological provenance concepts (PROV-O) for adequate answering. For simple PQs, following concepts apply: **When** can be mapped to `xsd:dateTime`, indicating a date, timestamp or duration. **Where** can be mapped to `prov:location`, e.g., physical place, path or list of ids. **Who** denotes the agents involved in the event, can be mapped to `prov:agent`. Complex PQs, however, have either uncovered concepts or require more complex answering: **How** describes the action leading up to the event, usually a (chain of) `prov:activity`. **Which** specifies the programs or instruments used in the event and can be linked to different concepts. **Why** represents the reasons for the event and needs additional reasoning or contextualisation of the entire system or parts of the system.

XAI-Specific Provenance Questions. Looking again in the *XAI Question Bank* [Liao *et al.*, 2021a] **what**-, **how**- or **why**-questions can be further specified to **what if**, **how to be** and **why not**. These will be answered by the correlation between forecast and the requested change (**what if**), feature highlights or ranges depending on the predictions (**how to be**), or changes that are required for alternative predictions (**why not**). To automate this process, we discuss next a reference architecture.

5 Conceptual Reference Architecture

After having provided the main definitions for PQs and their answers, classification and transformation steps are now illustrated in a conceptual reference architecture for our two-layered framework (cf. Figure 4). To support AI system transparency, it consists of an AI system and a provenance layer. While the AI system layer depicts an AI system lifecycle, the provenance layer supports provenance by design. The second layer includes four components: i) *question selection*, ii) *linguistic analysis of provenance questions*, iii) *mapping to provenance concepts*, and iv) *provenance data collection & analysis* (gray boxes in Figure 4).

- **Question selection.** In this component, users can upload a machine-readable AI system description and either loading applicable PQs from an existing provenance questionbank or providing custom PQs in natural language. Suitable technologies would be frontend frameworks coupled with natural language grammar frameworks such as *Grammatical Framework* [Ranta, 2004], but also recent generative AI approaches to support the translation of natural language questions. A questionbank based on existing works [Liao *et al.*, 2020; Naja and et al., 2022; European Commission, 2021] can be stored in relational or graph-based data, depending on use case needs.
- **Linguistic analysis of provenance questions.** The selected PQs or custom ones are analysed, using *Natural Language Processing (NLP)* and the W7 model. Suitable technologies are NLP libraries (e.g., SpaCy[2] or HuggingFace Transfomers[3], language models, as well as semantic web technologies libraries, e.g., Apache Jena[4], RDFLib[5]) or networkX for other graphs [6].
- **Mapping to provenance concepts.** This framework component processes the analysed question component to P-Plan and PROV-O concepts, other applicable ontologies or domain models. This is done to create provenance queries, mapping files, applicable trace templates for the needed data format (e.g., `JSON`). Suitable technologies are custom applications, either based on established semantic web tools such as or RDFLib to create and handle the question-to-query and query-to-trace-template mappings accordingly to selected ontologies and data models. However, also relational data can be used to guide the PQ formalisation.

[2]https://spacy.io
[3]https://huggingface.co/docs/transformers/en/index
[4]https://jena.apache.org
[5]https://rdflib.readthedocs.io/en/stable/
[6]https://networkx.org

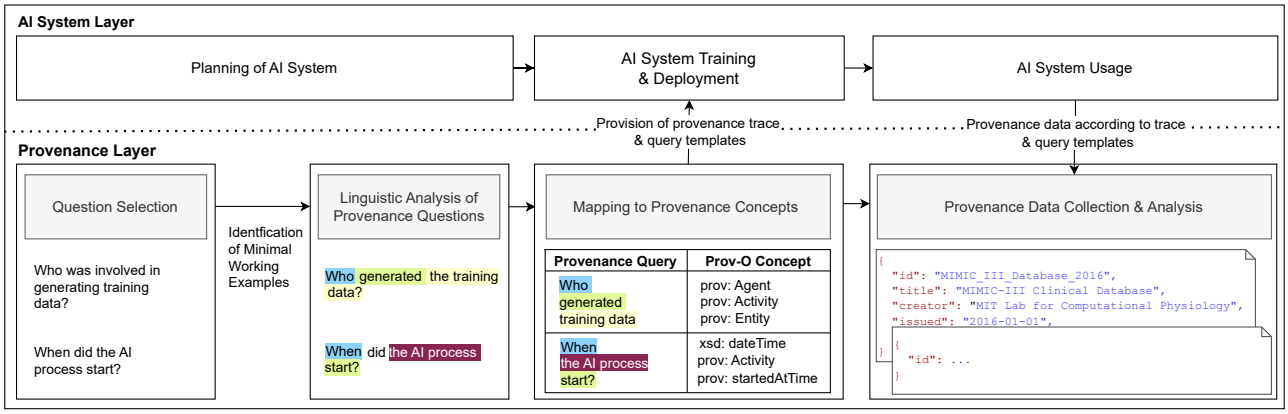

Figure 4: Conceptual reference architecture for the PQ framework.

- **Provenance data collection & analysis.** Finally, incoming trace data is integrated and stored for retrieval. Suitable technologies for this task are established graph-based databases, such as GraphDB[7] or Stardog[8] for knowledge graphs, as well as Neo4j[9] for property graphs. Also relational databases can be used.

Figure 4 illustrates how a sample run for two PQs could be done. First, the question is selected *Who was involved in generating the training data?* is selected, then linguistically analysed (question word, main verb and phrase) and mapped to provenance concepts (`prov:Agent`, `prov:Activity` and `prov:Entity`). Finally, the provenance data (different provenance traces) is collected and analysed.

6 Related Work

Provenance is an established concept in traditional information systems, closely linked to traceability, documentation and transparency [Herschel *et al.*, 2017]. There are methodologies to collect provenance in a structured way [Miles *et al.*, 2011]. Data models for modeling provenance are, for example, PROV-O [Lebo *et al.*, 2013] and the P-Plan ontology [Garijo and Gil, 2012]. Categorisations for traditional PQs have been proposed, for example, in [Zerva *et al.*, 2013]. Additionally, systems for generating provenance templates have been implemented to increase provenance trace quality, for instance [Curcin *et al.*, 2017]. However, a recent survey on biomedical data and workflows still indicates a gap in research about the completeness and quality of provenance data calling for automated solutions and specifications for provenance capturing [Gierend *et al.*, 2024]. Furthermore, provenance and traceability have gained importance in AI to achieve accountability and transparency [Turri and Dzombak, 2023; European Commission, 2021]. There are remaining gaps in operationalising governance and provenance for AI systems [Bhat *et al.*, 2023; Breit and et al., 2023].

Existing questionbanks, collecting natural language questions for different use cases, such as for *explainable AI (XAI)* [Liao *et al.*, 2020], for accountable AI [Naja and et al., 2022] or from documentation templates [Mitchell *et al.*, 2019; Gebru *et al.*, 2018], are valuable knowledge repositories for provenance aspects. However, they either remain high-level as the former example [Liao *et al.*, 2021a] or focus on specific provenance data models as the latter [Breit and et al., 2023]. These models also lack domain-specific tailoring and the ease of use for domain experts without provenance background. Moreover, AI system transparency considerations are always associated with a certain cost [Yoo, 2024] and thus, an implementation should be facilitated for different systems and stakeholders.

7 Conclusion

We have presented an approach for provenance question-based AI transparency and governance. For this, we first described an application scenario from explainable AI in the medical domain. To illustrate our approach, we demonstrated a provenance data model as well as provenance trace templates for two example PQs.

We introduced a conceptual reference architecture for a PQ framework consisting of four components. For the question selection, we defined simple and complex PQs as well as provenance results. We further detailed a linguistic analysis of the PQs and their mapping to provenance concepts. The aim of this framework is to provide support for deriving provenance requirements from PQs and to enable improved validation of provenance traces throughout the lifecycle of an AI system for different users and stakeholders. With this, we increase the overall transparency level of the AI system and the quality of provenance traces and logs.

In future work, we will implement the system and extend our concept towards complex PQs and the ones not qualifying our current definition, e.g., *Is [feature X] used or not used for predictions?* Furthermore, we will investigate the integration of other symbolic resources, e.g., ontologies [Vázquez-Flores *et al.*, 2022] for ethical issues or explanations [Chari *et al.*, 2020] to cover more question and answer types.

[7] https://graphdb.ontotext.com
[8] https://www.stardog.com
[9] https://neo4j.com

Acknowledgments

Dominique Haulser is funded by the German Research Foundation research grant 385808805; Laura Waltersdorfer is funded by the European Union's Horizon research grant 101120323.

References

[Auge *et al.*, 2024] Tanja Auge, Sascha Genehr, Meike Klettke, Frank Krüger, and Max Schröder. Towards dimensions and granularity in a unified workflow and data provenance framework. Presented at LWDA, 2024.

[Bhat *et al.*, 2023] Avinash Bhat, Austin Coursey, Grace Hu, Sixian Li, Nadia Nahar, Shurui Zhou, Christian Kästner, and Jin L. C. Guo. Aspirations and practice of ML model documentation: Moving the needle with nudging and traceability. In *CHI*, pages 749:1–749:17. ACM, 2023.

[Breit and et al., 2023] Anna Breit and et al. Combining Semantic Web and Machine Learning for Auditable Legal Key Element Extraction. In *ESWC*, pages 609–624. Springer, 2023.

[Breit *et al.*, 2023] Anna Breit, Laura Waltersdorfer, Fajar J. Ekaputra, Marta Sabou, Andreas Ekelhart, Andreea Iana, Heiko Paulheim, Jan Portisch, Artem Revenko, Annette ten Teije, and Frank van Harmelen. Combining machine learning and semantic web: A systematic mapping study. *ACM Comput. Surv.*, 55(14s):313:1–313:41, 2023.

[Chari *et al.*, 2020] Shruthi Chari, Oshani Seneviratne, Daniel M Gruen, Morgan A Foreman, Amar K Das, and Deborah L McGuinness. Explanation Ontology: A Model of Explanations for User-Centered AI. In *ISWC*, pages 228–243. Springer, 2020.

[Curcin *et al.*, 2017] Vasa Curcin, Elliot Fairweather, Roxana Danger, and Derek Corrigan. Templates as a Method for Implementing Data Provenance in Decision Support Systems. *Journal of biomedical informatics*, 65:1–21, 2017.

[European Commission, 2021] European Commission. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act), 2021. COM/2021/206 final.

[Fernandez *et al.*, 2023] Izaskun Fernandez, Cristina Aceta, Eduardo Gilabert, and Iker Esnaola-Gonzalez. Fides: An ontology-based approach for making machine learning systems accountable. *Journal of Web Semantics*, 79:100808, 2023.

[Garijo and Gil, 2012] Daniel Garijo and Yolanda Gil. Augmenting PROV with plans in P-PLAN: scientific processes as linked data. In *LISC@ISWC*, volume 951 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012.

[Gebru *et al.*, 2018] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *CoRR*, abs/1803.09010, 2018.

[Gierend *et al.*, 2024] Kerstin Gierend, Frank Krüger, Sascha Genehr, Francisca Hartmann, Fabian Siegel, Dagmar Waltemath, Thomas Ganslandt, and Atinkut Alamirrew Zeleke. Provenance information for biomedical data and workflows: Scoping review. *J Med Internet Res*, 26:e51297, Aug 2024.

[Herschel *et al.*, 2017] Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. A survey on provenance: What for? what form? what from? *VLDB J.*, 26(6):881–906, 2017.

[Johnson *et al.*, 2016] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[Lebo *et al.*, 2013] Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. Prov-o: The prov ontology. *W3C recommendation*, 30, 2013.

[Liao *et al.*, 2020] Q. Vera Liao, Daniel M. Gruen, and Sarah Miller. Questioning the AI: informing design practices for explainable AI user experiences. In *CHI*, pages 1–15. ACM, 2020.

[Liao *et al.*, 2021a] Q Vera Liao, Milena Pribić, Jaesik Han, Sarah Miller, and Daby Sow. Question-Driven Design Process for Explainable AI User Experiences. *arXiv preprint arXiv:2104.03483*, 2021.

[Liao *et al.*, 2021b] Q. Vera Liao, Moninder Singh, Yunfeng Zhang, and Rachel K. E. Bellamy. Introduction to explainable AI. In *CHI Extended Abstracts*, pages 127:1–127:3. ACM, 2021.

[Lundberg and Lee, 2017] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NIPS*, pages 4765–4774, 2017.

[Miles *et al.*, 2011] Simon Miles, Paul Groth, Steve Munroe, and Luc Moreau. Prime: A Methodology for Developing Provenance-Aware Applications. *ACM TOSEM*, 20(3):1–42, 2011.

[Mitchell *et al.*, 2019] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019.

[Naja and et al., 2022] Iman Naja and et al. Using Knowledge Graphs to Unlock Practical Collection, Integration, and Audit of AI Accountability Information. *IEEE Access*, 10:74383–74411, 2022.

[Netherlands Court of Audit, 2021] Netherlands Court of Audit. Understanding algorithms, January 2021. Accessed: 2024-11-22.

[Oppold and Herschel, 2020] Sarah Oppold and Melanie Herschel. Accountable data analytics start with accountable data: The liquid metadata model. In *ER Forum/-Posters/Demos*, volume 2716 of *CEUR Workshop Proceedings*, pages 59–72. CEUR-WS.org, 2020.

[Ram and Liu, 2009] Sudha Ram and Jun Liu. A new perspective on semantics of data provenance. In *SWPM*, volume 526 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009.

[Ranta, 2004] Aarne Ranta. Grammatical framework. *Journal of Functional Programming*, 14(2):145–189, 2004.

[Stearns *et al.*, 2001] Michael Q. Stearns, Colin Price, Kent A. Spackman, and Amy Y. Wang. SNOMED clinical terms: overview of the development process and project status. In *AMIA*. AMIA, 2001.

[Stromswold, 1995] Karin Stromswold. The Acquisition of Subject and Object Wh-Questions. *Language Acquisition*, 4:5–48, 1995.

[Turri and Dzombak, 2023] Violet Turri and Rachel Dzombak. Why we need to know more: Exploring the state of AI incident documentation practices. In *AIES*, pages 576–583. ACM, 2023.

[UNESCO, 2021] C UNESCO. Recommendation on the Ethics of Artificial Intelligence, 2021.

[Vázquez-Flores *et al.*, 2022] Karen Leticia Vázquez-Flores, Elena Montiel-Ponsoda, and María Poveda-Villalón. EA-Ontology: Ethical Assessment Ontology. *EKAW (Companion)*, 2022.

[Westergaard, 2009] Marit Westergaard. Usage-based vs. rule-based learning: the acquisition of word order in wh-questions in English and Norwegian. *Journal of Child Language*, 36(5):1023–1051, 2009.

[Wexler *et al.*, 2019] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The What-if Tool: Interactive Probing of Machine Learning Models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.

[Yoo, 2024] Christopher S. Yoo. Toward implementable ai standards. In *International Joint Conference on Artificial Intelligence 2024 Workshop on AI Governance: Alignment, Morality, and Law*, 2024.

[Zerva *et al.*, 2013] Paraskevi Zerva, Steffen Zschaler, and Simon Miles. Towards design support for provenance awareness: a classification of provenance questions. In *EDBT/ICDT Workshops*, pages 275–281. ACM, 2013.