



AUS DER FORSCHUNG

Was ist Datenprovenienz? Der Kontext hinter einer Zahl

Datenprovenienz dokumentiert Entitäten, Aktivitäten und Akteure hinter einem Ergebnis. Ein Herkunftspfad macht einen Anspruch prüfbar und reproduzierbar.

Sinan Isoglu, MBA (Quantic)

Führungskraft für Umsatzwachstum, Dozent und Doktorand

VERÖFFENTLICHT

6. September 2026

LESEZEIT

5 Min.

WÖRTER

992

<https://isoglu.com/de/insights/journal/was-ist-datenprovenienz/>

Management Summary	2
Was bedeutet Datenprovenienz?	3
Welche Felder machen Provenienz prüfbar?	3
Wie sieht ein Provenienzregister aus?	4
Wo kann Provenienz scheitern?	5
Wie prüft ein Team einen Provenienzdatensatz?	6
Quellen	7

MANAGEMENT SUMMARY

Datenprovenienz dokumentiert den Kontext, in dem ein Datenobjekt erzeugt, verändert, zugeschrieben und verwendet wurde. Das Modell PROV-DM des W3C liefert dafür eine begrenzte Sprache aus Entitäten, Aktivitäten, Akteuren und Ableitungsbeziehungen. Dieser Artikel trennt Provenienz von Daten-Lineage, Quellenprüfung, Datenqualität und Ereigniserfassung. Ein synthetisches Provenienzregister zeigt, wie Quellenidentität, Transformation, Verantwortung, Zeit und spätere Nutzung sichtbar bleiben, ohne aus einer dokumentierten Spur die Richtigkeit der Quelle oder Kausalität des Ergebnisses abzuleiten.

Schlagwörter: Datenprovenienz · Daten-Lineage · Quellenprüfung · Reproduzierbarkeit · Claim-Provenienz · Prüfpfad

Ein Dashboard kann die fertige Zahl bewahren und zugleich den Kontext verlieren, der sie prüfbar macht. Die Quelle wurde von jemandem heruntergeladen. Ein Filter wurde in einem Notebook geändert. Eine Führungskraft genehmigte einen Ausschluss. Im Bericht steht weiterhin „qualifizierte Accounts“, aber eine andere Person kann nicht erkennen, welches Objekt verwendet wurde oder wer es veränderte.

Datenprovenienz ist der Datensatz der Entitäten, Aktivitäten und Akteure, die an der Erzeugung, Beeinflussung oder Bereitstellung eines Datenobjekts beteiligt waren. Sie gibt einem Ergebnis Kontext. Sie bescheinigt weder die Richtigkeit der Quelle noch die Gültigkeit der Transformation oder die Kausalität des Schlusses.

Der Artikel zu Daten-Lineage behandelt den angewandten Pfad von Eingaben über Joins, Filter und Transformationen. Diese Seite behandelt den weiteren Kontext dieses Pfads: Welche Objekte existierten, welche Aktivität sie verband, wer dafür verantwortlich war und welche spätere Nutzung vom Ergebnis abhing.

Was bedeutet Datenprovenienz?

Das PROV-DM-Modell des World Wide Web Consortium beschreibt Provenienz über Entitäten, Aktivitäten und Akteure. Es enthält Beziehungen für Nutzung, Erzeugung, Ableitung, Zuordnung und Zuschreibung (World Wide Web Consortium, 2013). Diese Begriffe sind nützlich, weil „Das Dashboard zeigte es“ noch kein Prüfpfad ist.

Die benachbarten Objekte bleiben getrennt:

Tabelle 1 Was bedeutet Datenprovenienz?

Objekt	Was es dokumentiert	Was es nicht feststellt
Provenienz	Entitäten, Aktivitäten, Akteure, Beziehungen und Zeit	Dass jede Quelle oder Operation richtig ist
Daten-Lineage	Den angewandten Weg von einer Eingabeversion zu Feld oder Ergebnis	Dass das Maß das gewünschte Konstrukt abbildet
Quellenprüfung	Ob ein zitiertes Dokument oder ein Datensatz identifizierbar vorliegt	Dass die Quelle jeden daraus gezogenen Satz trägt
Datenqualität	Ob ein Feld oder Ergebnis eine erklärte Eignungsregel besteht	Dass das Ergebnis kausal ist
Ereigniserfassung	Was wann mit welcher Einheit geschah	Warum das Ereignis geschah

Tabelle aus diesem Essay. Quellen und Einordnung stehen im Beitrag.

Provenienz ist damit ein Kontextobjekt. Lineage ist ein Transformationspfad. Quellenprüfung ist ein Evidenzurteil. Qualität ist ein Eignungsurteil. Ereigniserfassung ist ein Beobachtungsdatensatz.

Welche Felder machen Provenienz prüfbar?

Beginnen Sie mit dem Objekt, dem eine lesende Person vertrauen soll. Halten Sie anschließend die Felder fest, die seinen Kontext rekonstruierbar machen:

- 1 Entität: Quelldatei, Datensatz, Tabelle, Bericht, Modellergebnis oder Entscheidungsobjekt.
- 2 Identität und Version: stabile Kennung, Abrufzeit, Prüfsumme oder Snapshot-Bezeichnung.

- 3 Aktivität: Extraktion, Normierung, Join, Aggregation, manuelle Prüfung oder Veröffentlichung.
- 4 Akteur: Person, Team, System, Anbieter oder Prozess, der die Aktivität verantwortet.
- 5 Beziehung: genutzt, erzeugt, abgeleitet, zugeschrieben, zugeordnet, ersetzt oder entwertet.
- 6 Zeit: Ereignis-, Abruf-, Verarbeitungs- und Entscheidungszeit, wenn sie auseinanderfallen.
- 7 Nutzung: Dashboard, Modell, Claim, Meeting, Kundenaktion oder Regel, die das Objekt nutzte.
- 8 Ausnahme: fehlende Quelle, geänderte Regel, verspätete Daten, fehlender Test oder offene Verantwortung.

Die Identität ist entscheidend. Eine Datei namens customers.csv ist kein dauerhaftes Objekt, wenn ihr Inhalt jeden Morgen wechselt. Ein versionierter Snapshot oder ein Inhalts-Hash unterscheidet das Objekt von seinem Nachfolger. Auch das Akteursfeld ist wichtig. „Automatisierung“ kann verbergen, ob ein Anbieter, ein geplanter Job oder eine Person die relevante Änderung vornahm.

Wie sieht ein Provenienzregister aus?

Die sechs Zeilen sind synthetisch. Sie enthalten keine privaten Quellen, Teilnehmenden, Kunden- oder Unternehmensdaten. Sie zeigen, wie dasselbe Ergebnis sowohl einen Transformationspfad als auch einen Kontextdatensatz tragen kann.

Abbildung 1 Das synthetische Provenienzregister

ID	Entität	Aktivität	Akteur	Beziehung und Zeit	Nachgelagerte Nutzung
P-01	Quell-PDF S-17, abgerufen 2026-08-31	Tabelle 4 extrahieren	Research Analyst	Bei Abruf genutzt	Claim-Ledger
P-02	Extrakt E-17, Version 1	Einheiten normieren	Analyse-Note book N-4	Aus S-17 abgeleitet	Arbeitsdatensatz
P-03	Account-Snapshot A-31	Join über Account-ID	Geplante Pipeline	Am 2026-09-01 06:00 zugeordnet	Kohortentabelle
P-04	Kohortentabelle C-2	Doppelte IDs ausschließen	Data Steward	Aktivität protokolliert; Grund bleibt erhalten	Prüfblatt
P-05	Prüfblatt W-9	Evidenzsatz freigeben	Redaktion und Quellenverantwortung	Am 2026-09-02 W-9 zugeschrieben	Artikelentwurf
P-06	Artikelentwurf D-1	Veröffentlichen oder halten	Redaktionsprozess	Halten bei offener Quelle oder Beziehung	Freigabeentscheidung

Synthetisches Provenienzregister des Autors auf Grundlage von PROV-DM des World Wide Web Consortium (2013). Das Vokabular ist quellengebunden; alle Entitäten, Aktivitäten, Akteure, Zeitpunkte und Dispositionen sind illustrativ.

P-01 benennt die gehaltene Entität. P-02 dokumentiert eine Operation, statt des Extrakts stillschweigend als Originaldaten zu behandeln. P-05 bewahrt die Zuschreibung des Evidenzsatzes. P-06 macht die Freigabeentscheidung selbst zum Datensatz. Wenn P-01 nicht identifizierbar ist oder P-04 keine Doppelregel besitzt, kann das Ergebnis trotz plausibler Zahl gehalten werden.

Wo kann Provenienz scheitern?

Provenienz kann vollständig sein und trotzdem einen ungültigen Prozess sichtbar machen. Häufige Fehler sind:

- Unversionierter Austausch: Eine benannte Datei verändert sich, während ihr Pfad gleich bleibt.
- Verlust der Beziehung: Eine abgeleitete Tabelle bleibt, aber die erzeugende Operation verschwindet.
- Zusammengelegter Akteur: „Das System“ verbirgt eine manuelle Änderung, einen Anbieterwechsel oder einen neuen Eigentümer.
- Verwechselte Zeit: Abrufzeit wird als Ereignis- oder Entscheidungszeit behandelt.
- Unprotokollierter Ausschluss: Eine Zeile wird ohne Regel, Grund oder Prüfung entfernt.
- Überdehnung der Zitation: Eine gehaltene Quelle wird so verwendet, als trüge sie eine stärkere Aussage.

Das W3C-Modell unterstützt interoperable Beschreibung und Schlussfolgerung, bescheinigt aber nicht, dass Quelle oder Transformation korrekt sind. Deshalb muss Provenienz neben Qualität, Validität und Kausalprüfung stehen, nicht an deren Stelle.

Wie prüft ein Team einen Provenienzdatensatz?

Wählen Sie ein Ergebnis und gehen Sie von seiner Nutzung rückwärts:

- 1 Benennen Sie Entität, Einheit, Version und zu prüfende Entscheidung.
- 2 Lösen Sie das Quellobjekt und die Abruf- oder Snapshot-Grenze auf.
- 3 Listen Sie jede Aktivität in der richtigen Reihenfolge, einschließlich manueller oder externer Schritte.
- 4 Benennen Sie Akteur und Zeitpunkt für jede Aktivität.
- 5 Prüfen Sie die Beziehungen: genutzt, erzeugt, abgeleitet, zugeschrieben, ersetzt oder entwertet.
- 6 Ordnen Sie jeden Evidenzsatz genau dem Quellobjekt zu, das ihn tatsächlich trägt.
- 7 Markieren Sie das Ergebnis als reproduzierbar, gehalten, neu ausgestellt oder nicht rekonstruierbar.

Provenienz macht Kontext sichtbar. Sie macht eine schwache Quelle nicht stark, ein Maß nicht valide und ein deskriptives Ergebnis nicht kausal.

Der Artikel zur Reproduzierbarkeit stellt die nächste Frage: ob eine andere Person den erklärten Pfad rekonstruieren kann.

World Wide Web Consortium. (2013). PROV-DM: The PROV Data Model. W3C Recommendation. Quellenseite

ENDE DES BEITRAGS

ZITIERWEISE

Isoglu, S. (2026, 6. September). Was ist Datenprovenienz? Der Kontext hinter einer Zahl.
Sinan Isoglu. <https://isoglu.com/de/insights/journal/was-ist-datenprovenienz/>

WEITERLESEN

Aus der Forschung

Was ist Data Lineage? Die Transformationen hinter einem Ergebnis

<https://isoglu.com/de/insights/journal/was-ist-data-lineage/>

Aus der Forschung

Was ist Quellenprüfung? Von der Aussage zur Version of Record

<https://isoglu.com/de/insights/journal/was-ist-quellenpruefung/>

Aus der Forschung

Was ist Reproduzierbarkeit? Kann eine andere Person das Ergebnis rekonstruieren?

<https://isoglu.com/de/insights/journal/was-ist-reproduzierbarkeit/>



Sinan Isoglu, MBA (Quantic)

Führungskraft für Umsatzwachstum, Dozent und Doktorand