Docket

DOCKET / RESEARCH & PRACTICAL GUIDANCE

Data Provenance for Deal Research: Using the W3C PROV Model

> Build a more reviewable acquisition pipeline with Docket. Our deal-origination software and managed research help private equity teams screen companies against a mandate and understand the evidence behind each finding.

Published by Docket · 2026-10-29 · data provenance · provenance meaning · data lineage research · evidence trail

Original article: https://docket.capital/articles/data-provenance-w3c-prov-deal-research



In this article

1. What provenance means
2. The PROV core model
3. From PROV to a company fact
4. A minimal provenance record
5. Hashing and archiving the source
6. Auditing a provenance trail
7. Standards map: PROV, Dublin Core, WARC, FIPS 180-4
8. What provenance cannot do
9. Sources

What provenance means

A research file on a private company is a pile of claims: headcount, locations, ownership, services, licences. A reviewer who opens that file asks the same questions of every claim. Where did it come from? When was it read? Who or what extracted it? Has the source changed since?

The W3C's PROV data model gives those questions a formal shape. It defines provenance as "information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness" (W3C PROV-DM). The first half is a record of production; the second is the purpose, assessment. Provenance does not make a claim true. It gives the reviewer what they need to judge it.

Many deal-research errors are errors of lineage: a headcount copied from a vendor record that copied an old press release, or a service line read from a similarly named company's site. Each is cheap to catch if the file records how the value was produced.

This article covers the open standard. For the review principles that sit on top of a provenance record (source ranking, quotation versus conclusion, retained copies), see Docket's evidence standards.

The PROV core model

PROV is a family of W3C documents dated 30 April 2013. The data model (PROV-DM), the ontology (PROV-O), the notation (PROV-N) and PROV-CONSTRAINTS are Recommendations; the primer, an XML schema, a Dublin Core mapping and others are Working Group Notes (W3C PROV Overview). Start with the data model and the PROV Model Primer, which walks through the concepts with examples.

The model rests on three types (W3C PROV-DM):

PROV type	W3C definition (abridged)	In target research
Entity	"a physical, digital, conceptual, or other kind of thing with some fixed aspects"	A live web page, a saved copy of it, a quoted passage, the recorded value
Activity	"something that occurs over a period of time and acts upon or with entities"	Retrieving a page, extracting a value, reviewing an answer
Agent	"something that bears some form of responsibility for an activity taking place, for the existence of an entity, or for another agent's activity"	An analyst, a crawler or extraction model, the firm the analyst works for

The "fixed aspects" phrase matters: the page as you retrieved it is a different entity from the page today, and PROV lets you say which one your claim rests on.

Seven core relations connect the three types (W3C PROV-DM):

Relation	Connects	Definition (abridged)
wasGeneratedBy	entity → activity	Generation is "the completion of production of a new entity by an activity"

Relation	Connects	Definition (abridged)
used	activity → entity	Usage is "the beginning of utilizing an entity by an activity"
wasDerivedFrom	entity → entity	A derivation is "a transformation of an entity into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity"
wasAttributedTo	entity → agent	Attribution is "the ascribing of an entity to an agent"
wasAssociatedWith	activity → agent	Association is "an assignment of responsibility to an agent for an activity"
actedOnBehalfOf	agent → agent	Delegation: an agent carries out an activity for another agent, which "retains some responsibility for the outcome"
wasInformedBy	activity → activity	Communication: one activity uses an entity generated by another

PROV-DM also defines specialised kinds of derivation. Two are directly useful in diligence. A **quotation** is "the repeat of (some or all of) an entity, such as text or image, by someone who may or may not be its original author", and a **primary source** is "something produced by some agent with direct experience and knowledge about the topic, at the time of the topic's study, without benefit from hindsight" (W3C PROV-O). Quotation ties a value to the exact sentence it came from; primary source marks that an ownership claim came from a state filing rather than a directory summarising it.

A **bundle** is "a named set of provenance descriptions, and is itself an entity" (W3C PROV-DM), so a batch's provenance record can itself be attributed and dated.

From PROV to a company fact

Illustrative example (fictional company, fictional values). An analyst screening commercial HVAC contractors records that Harborline Mechanical has about 40 employees. The value came from the company's About page, which reads "Our team of 40 technicians and staff serves the tri-county area." Mapped onto PROV, the record contains:

1. **Entity: the live page** at `https://harborline.example/about`. It exists independently of the research and can change at any time.
2. **Activity: retrieval.** A crawler fetched the page on 20 October 2026 at 14:03 UTC. The activity *used* the live page and is *associated with* the crawler (a software agent).
3. **Entity: the capture.** The saved copy of the page *was generated by* the retrieval and *was derived from* the live page. Unlike the live page, it has fixed content and can be hashed.
4. **Entity: the quotation.** The sentence "Our team of 40 technicians and staff…" *was quoted from* the capture.
5. **Activity: extraction.** The analyst read the capture and recorded a value. The activity *used* the capture and *was associated with* the analyst, who *acted on behalf of* the fund.

6. **Entity: the claim.** `employee_count = 40` *was generated by* extraction, *was derived from* the quotation, and *was attributed to* the analyst.

Two things fall out. The claim is about a statement on a page, not about payroll. And every hop can be checked independently: a reviewer can open the capture, find the quotation and judge whether a marketing sentence supports "40 employees".

Here is the same record in PROV-O, the RDF encoding of PROV (W3C PROV-O). The `ex:` namespace, identifiers and the hash value are placeholders.

```
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix xsd:  <http://www.w3.org/2001/XMLSchema#> .
@prefix ex:   <https://research.example/ns#> .

# Illustrative: fictional company, placeholder identifiers and hash
<https://harborline.example/about> a prov:Entity .

ex:retrieval-0412 a prov:Activity ;
    prov:used <https://harborline.example/about> ;
    prov:startedAtTime "2026-10-20T14:03:11Z"^^xsd:dateTime ;
    prov:endedAtTime   "2026-10-20T14:03:12Z"^^xsd:dateTime ;
    prov:wasAssociatedWith ex:crawler .

ex:capture-0412 a prov:Entity ;
    prov:wasGeneratedBy ex:retrieval-0412 ;
    prov:wasDerivedFrom <https://harborline.example/about> ;
    prov:generatedAtTime "2026-10-20T14:03:12Z"^^xsd:dateTime ;
    ex:sha256 "PLACEHOLDER-NOT-A-REAL-DIGEST" .

ex:quote-0412 a prov:Entity ;
    prov:value "Our team of 40 technicians and staff serves the tri-county area." ;
    prov:wasQuotedFrom ex:capture-0412 .

ex:extraction-0412 a prov:Activity ;
    prov:used ex:capture-0412 ;
    prov:wasAssociatedWith ex:analyst-17 .

ex:claim-0412 a prov:Entity ;
    prov:value 40 ;
    ex:field "employee_count" ;
    prov:wasGeneratedBy ex:extraction-0412 ;
    prov:wasDerivedFrom ex:quote-0412 ;
    prov:wasAttributedTo ex:analyst-17 .

ex:crawler    a prov:SoftwareAgent .
ex:analyst-17 a prov:Person ; prov:actedOnBehalfOf ex:fund .
ex:fund       a prov:Organization .
```

Every `prov:` term above is defined in PROV-O, either in its "Starting Point" set or among its expanded terms (W3C PROV-O). `ex:sha256` and `ex:field` are local extensions, which PROV-O allows. The shape also fits in a spreadsheet.

A minimal provenance record

The flat equivalent for a CRM or spreadsheet: one row per claim, each column mapped to its PROV concept.

Field	Example value (illustrative)	PROV mapping	Why a reviewer needs it
subject_entity	Harborline Mechanical (registry ID if resolved)	Attribute of the claim	Catches wrong-company sourcing
field	employee_count	Attribute of the claim	Ties the value to a defined question
value	40	`prov:value`	The answer itself
quote	"Our team of 40 technicians and staff…"	Entity, `wasQuotedFrom` capture	Shows the exact support
source_url	https://harborline.example/about	Entity used by retrieval	Where the material lived
source_type	company website (self-reported)	`prov:type`, or primary-source flag	Separates first-hand from summary
retrieved_at	2026-10-20T14:03:12Z	`generatedAtTime` of the capture	Dates the evidence, not the research
retrieval_method	scripted fetch; WARC	Activity plus associated software agent	Explains how the copy was made
capture_ref	warc/2026-10-20/0412.warc.gz	Entity generated by retrieval	Locates the retained copy
capture_sha256	(64 hex characters)	Attribute of the capture	Detects later change to the copy
extracted_by	analyst-17	`wasAssociatedWith` / `wasAttributedTo`	Accountability
reviewed_by, reviewed_at	reviewer-04, 2026-10-22	Second activity using the claim	Records that review happened
status	accepted / open / struck	Local extension	Keeps unresolved claims visible

Three design rules:

- **Separate the retrieval date from the research date.** A claim recorded today from a capture made six months ago is six months old.
- **Store the quote, not just the URL.** A live page can change after you rely on it; the reviewer needs what you actually read.

- **Never overwrite a claim.** A corrected value is a new entity derived from the old one (PROV-DM's revision: "a derivation for which the resulting entity is a revised version of some original" (W3C PROV-DM)). Overwriting destroys the trail.

If your team uses Dublin Core, its DCMI Metadata Terms overlap with this record. They include `source`, defined as "a related resource from which the described resource is derived", `created`, and `provenance`, defined as "a statement of any changes in ownership and custody of the resource since its creation that are significant for its authenticity, integrity, and interpretation", with a `ProvenanceStatement` class (DCMI Metadata Terms). Dublin Core's "provenance" concerns custody, narrower than PROV's sense; the W3C's PROV-DC Note gives a partial mapping between the two (W3C PROV-DC).

Hashing and archiving the source

A capture is evidence only if you can show it has not changed. NIST's Secure Hash Standard, FIPS 180-4, "specifies hash algorithms that can be used to generate digests of messages", and "the digests are used to detect whether messages have been changed since the digests were generated" (NIST FIPS 180-4). SHA-256 is one of the algorithms it specifies (FIPS 180-4 PDF). On 28 October 2026 the CSRC page showed the August 2015 version, with a March 2023 planning note that NIST "has decided to revise FIPS 180-4"; no successor (FIPS 180-5) had been published, so FIPS 180-4 remains the current standard.

Rules for hashing captures:

1. **Hash at capture time, over the exact bytes you store.** A later hash proves only that the copy is unchanged since that later moment.
2. **Record the algorithm with the value** (`sha256:` plus the hex digest). An unlabelled digest becomes ambiguous.
3. **Re-hash on audit.** A mismatch means the copy was altered, corrupted or replaced; every claim derived from it needs review.
4. **Do not over-read a match.** It shows the copy is unchanged, not that the content was true or that the page belonged to the company you think it did.

The established container format is WARC. The Library of Congress says it "specifies a method for combining multiple digital resources into an aggregate archival file together with related information", and that it was approved as an international standard (ISO 28500) in May 2009 (Library of Congress, WARC format description). ISO has since replaced the 2009 edition with ISO 28500:2017 (edition 2, published August 2017), which ISO lists as last reviewed and confirmed in 2023 (ISO 28500:2017). The IIPC's WARC 1.1 specification requires each record to carry a `WARC-Date`, defined as the instant data capture for the record began, and allows optional `WARC-Block-Digest` and `WARC-Payload-Digest` fields carrying a labelled digest; it states that "no particular algorithm is recommended" (IIPC, The WARC Format 1.1). If you store WARC files, record which digest algorithm your tool writes, and add your own SHA-256 if it writes something weaker.

A PDF print or screenshot is weaker but better than a bare URL. For pages that have already changed, a public web archive capture can stand in; Docket's Wayback Machine evidence protocol covers how to cite one.

Auditing a provenance trail

Audit procedure (template):

1. **Draw a sample** of claims from the batch. Size it for the defect rate you need to detect; the audit sampling worksheet gives zero-defect sample sizes and acceptance rules.
2. **Resolve each claim to its capture** using `capture_ref`. Missing capture: fail.
3. **Re-hash the capture** and compare with `capture_sha256`. Mismatch: fail and quarantine every claim derived from that capture.
4. **Find the quote in the capture.** Not present verbatim: fail.
5. **Check subject identity.** Does the capture belong to the entity in `subject_entity`? Wrong company: fail.
6. **Judge support.** Does the quote support the value as defined by `field`? Unsupported or over-read: fail.
7. **Check dates.** Is `retrieved_at` recent enough for the field's freshness rule? Stale: flag, not necessarily fail.
8. **Record the audit as its own activity**, with auditor, date and outcome, so the audit itself has provenance.

Illustrative tally (fictional): 50 sampled claims return 46 passes, 2 quotes missing from the capture, 1 wrong-company source and 1 stale flag. Three are defects, judged against the acceptance number set before sampling. Steps 2 to 4 can be scripted; steps 5 and 6 are judgement for someone other than the extractor.

Standards map: PROV, Dublin Core, WARC, FIPS 180-4

Standard	What it covers	What it leaves to you	Status (as checked 2026-10-28)
W3C PROV-DM / PROV-O	Concepts and RDF terms for entities, activities, agents and their relations; derivation, quotation, primary source, revision, bundles	What counts as a valid source; field definitions; review rules	W3C Recommendations, 30 April 2013 (PROV Overview)
DCMI Metadata Terms	Descriptive metadata: source, created, provenance statement and more	Process detail (who extracted what, how)	Issued 2020-01-20 per the DCMI page (DCMI)
WARC (ISO 28500; IIPC 1.1)	Container for captured web resources with capture time and optional digests	Which pages to capture; linking captures to claims	ISO 28500:2009, replaced by ISO 28500:2017 (confirmed 2023); IIPC publishes 1.0 and 1.1 (ISO; LoC)
NIST FIPS 180-4	Hash algorithms (including SHA-256) to detect change	When to hash; where to store digests	Final, August 2015; revision announced 2023, no successor published as of October 2026 (NIST)

What provenance cannot do

Provenance records lineage, not truth: it is information that "can be used to form assessments" (W3C PROV-DM). The assessment is still yours.

- **A perfect trail to a wrong source is still wrong.** If the company's own page overstates its headcount, the record faithfully documents an overstatement.
- **Provenance does not rank sources.** Preferring a filing over a self-published page is a policy question.
- **Provenance does not resolve conflicts.** When two well-documented sources disagree, the record shows both. Settling the disagreement needs a rule; Docket's conflicting-sources claim ledger sets out one approach.
- **Identity errors are upstream.** A capture of the wrong company's site produces an internally consistent, entirely wrong trail. Entity resolution has to happen before extraction, not after.
- **Granularity has a cost.** Record at the level where someone will ask "where did this come from?", usually the claim.

Docket's research follows the same logic: Scout's answers each quote a retained page, and Audit re-reads that material with fetch tools denied.

Sources

1. PROV-DM: The PROV Data Model, W3C Recommendation — W3C (2013), https://www.w3.org/TR/prov-dm/, accessed 2026-10-28
2. PROV-O: The PROV Ontology, W3C Recommendation — W3C (2013), https://www.w3.org/TR/prov-o/, accessed 2026-10-28
3. PROV-Overview: An Overview of the PROV Family of Documents, W3C Working Group Note — W3C (2013), https://www.w3.org/TR/prov-overview/, accessed 2026-10-28
4. PROV Model Primer, W3C Working Group Note — W3C (2013), https://www.w3.org/TR/prov-primer/, accessed 2026-10-28
5. Dublin Core to PROV Mapping (PROV-DC), W3C Working Group Note — W3C (2013), https://www.w3.org/TR/prov-dc/, accessed 2026-10-28
6. DCMI Metadata Terms — Dublin Core Metadata Initiative (2020), https://www.dublincore.org/specifications/dublin-core/dcmi-terms/, accessed 2026-10-28
7. WARC, Web ARChive file format (fdd000236) — Library of Congress, Sustainability of Digital Formats (2024), https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml, accessed 2026-10-28
8. The WARC Format 1.1 — International Internet Preservation Consortium (n.d.), https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/, accessed 2026-10-28
9. FIPS 180-4, Secure Hash Standard (SHS), publication page — NIST (2015), https://csrc.nist.gov/pubs/fips/180-4/upd1/final, accessed 2026-10-28
10. FIPS PUB 180-4, Secure Hash Standard (PDF) — NIST (2015), https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf, accessed 2026-10-28
11. ISO 28500:2017, Information and documentation — WARC file format — International Organization for Standardization (2017), https://www.iso.org/standard/68004.html, accessed 2026-10-28

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://www.w3.org/TR/prov-dm
2. https://www.w3.org/TR/prov-overview
3. https://www.w3.org/TR/prov-primer
4. https://www.w3.org/TR/prov-o
5. https://harborline.example/about
6. https://www.dublincore.org/specifications/dublin-core/dcmi-terms
7. https://www.w3.org/TR/prov-dc
8. https://csrc.nist.gov/pubs/fips/180-4/upd1/final
9. https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf
10. https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml
11. https://www.iso.org/standard/68004.html
12. https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1

THE PUBLISHER

About Docket

Docket provides deal-origination research software and managed research for private equity firms. We help investment teams investigate acquisition targets using structured screening criteria, retained sources and reviewable company evidence. Teams can work through a self-serve platform or use managed research, depending on how they want research delivered.

Research against a defined mandate

A useful target list needs more than company names. Docket focuses on the questions that determine whether a company fits an investment mandate, including the evidence needed to support or qualify each answer. Our research approach makes the connection between screening criteria, source material and conclusions visible to the team reviewing the work.

Triage, Scout and Audit

Docket's three named research agents perform complementary tasks. Triage resolves company identity and screens fit. Scout collects sourced answers against the mandate. Audit checks retained evidence, addresses contradictions and leaves unsupported answers visibly unresolved. This structure helps reviewers distinguish established findings from missing information and questions requiring further investigation.

Evidence that supports investment-team judgment

Our research library covers market mapping, screening criteria, private-company data, succession and ownership, source evaluation and evidence standards. These resources explain the methods and limitations behind origination research. Findings support a team's judgment; they do not establish that a company is for sale or guarantee a transaction or investment outcome.

Work with Docket

Visit Docket to explore the platform, managed research and the current contact path. Read about screening criteria, evidence standards, AI research agents and market mapping.

Public examples are illustrative unless explicitly identified otherwise. Research preparation and authorized outreach are separate activities; confidential target lists and customer outcomes should never be inferred from an educational example.

Website: https://docket.capital

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.