Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Anchoring a Claim to a Passage: Web Annotation Selectors

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-10-29 · web annotation textquoteselector · evidence span citation · anchor citation to text passage · text fragment selector · sentence level citation

Original article: https://reliablecontext.com/articles/anchoring-claims-to-passages-web-annotation-selectors



In this article

1. Why a URL is not a citation
2. The Web Annotation model in brief
3. TextQuoteSelector: anchoring by content
4. TextPositionSelector: anchoring by offset
5. Fragment, range and refinement
6. How each selector survives document change
7. Retained copies, states and hashes
8. A worked example
9. PDF and HTML differences
10. Measuring whether citations support claims
11. A verification routine
12. Summary
13. Sources

Why a URL is not a citation

A citation that supports a specific claim has to identify three things: the resource, the version of it that was read, and the passage within it that supports the claim. A bare URL identifies only the first, and only for as long as the address resolves to the same content. When the page is revised, the URL still resolves, the citation still looks valid, and the passage it was meant to point to may no longer exist.

The passage is what the site's glossary calls an evidence span: the exact stretch of source text that supports a claim. This article describes the standard way to encode an evidence span so that software can find it again and a reviewer can check it: the selectors and states of the W3C Web Annotation Data Model. It explains how each selector behaves when the document changes, why they should be combined, and how a retained copy and a content hash close the remaining gaps.

The Web Annotation model in brief

The Web Annotation Data Model is a W3C Recommendation dated 23 February 2017. An annotation links a *body* (here, the claim or a note about it) to a *target* (the source). When the target is part of a resource rather than all of it, the model uses a *Specific Resource* with three relevant components (section 4):

Component	Question it answers	Examples
`source`	Which resource?	The document's IRI
`state`	Which representation of it?	`TimeState` with `sourceDate`, `cached` copy
`selector`	Which part of that representation?	`TextQuoteSelector`, `TextPositionSelector`, `FragmentSelector`, `RangeSelector`

The specification defines a selector as the means "to describe how to determine the Segment from within the Source resource", and it allows several: "Multiple Selectors can be given to describe the same Segment in different ways in order to maximize the chances that it will be discoverable later" (section 4.2). Order of processing matters: "States MUST be processed before processing Selector or Style information" (section 4.3). In other words, first find the right version, then find the passage in it.

TextQuoteSelector: anchoring by content

A Text Quote Selector "describes a range of text by copying it, and including some of the text immediately before (a prefix) and after (a suffix) it to distinguish between multiple copies of the same sequence of characters" (section 4.2.4). It has three properties:

- `exact`: "A copy of the text which is being selected, after normalization." Exactly one is required.
- `prefix`: text immediately before; at most one, and one SHOULD be given.
- `suffix`: text immediately after; at most one, and one SHOULD be given.

Three rules from the specification determine whether two systems will agree on a match:

1. Selection is "in terms of unicode code points", not code units, and should not split a grapheme cluster.
2. Text "MUST be normalized before recording": HTML or XML tags should be removed and character entities replaced by the characters they encode.
3. If the prefix, exact text and suffix still match more than once, "the selection SHOULD be treated as matching all of the matches."

The strength of the quote selector is that it survives edits elsewhere in the document. Insert a paragraph above the passage and the quote is still found. Its weakness is the mirror image: if the passage itself is edited, the quote no longer matches, which is usually the right outcome for evidence, because the support has changed.

The specification also flags a rights issue. Because the selector copies text, it "is potentially dangerous" for content under copyright or distribution restrictions, and for such texts "the use of the Text Position Selector is perhaps more appropriate" (section 4.2.4, note).

TextPositionSelector: anchoring by offset

A Text Position Selector "describes a range of text by recording the start and end positions of the selection in the stream" (section 4.2.5). Position 0 is before the first character; `start` is included and `end` is not. In the specification's example, in "abcdefghijklmnopqrstuvwxyz", start 4 and end 7 select "efg". The same normalisation as for quotes must be applied before counting.

Offsets copy no text, which is why they suit restricted content. They are also, in the specification's words, "very brittle with regards to changes to the resource. Any edits or dynamically transcluded content may change the selection, and thus it is RECOMMENDED that a State be additionally used to help identify the correct representation." A single inserted word before the passage shifts every offset after it, and the selector silently points at different text.

Fragment, range and refinement

FragmentSelector reuses the fragment syntax a media type already defines, with `conformsTo` naming the specification (section 4.2.1). The specification lists, among others, plain text per RFC 5147 (example `char=0,10`) and PDF per RFC 3778. RFC 3778 has since been obsoleted by RFC 8118, which states that the PDF fragment specification "is now included in ISO 32000-2" and summarises parameters such as `page=<pageNum>` (first page is 1), `nameddest`, `structelem` and `comment`. For PDFs, `page=` is the most portable anchor; it locates a page, not a sentence, so it is a coarse first step.

RFC 5147 is notable for building in change detection: it lets plain-text fragment identifiers carry integrity-check information, either `length=` or `md5=`, "used to detect changes in the resource", so that applications "can then warn users about the possibility that a fragment identifier might have been broken" (RFC 5147, section 2.3).

RangeSelector uses two other selectors to mark a start and an end, for selections that cross internal boundaries, such as table cells; the selection runs from the beginning of the start selector "through to the beginning of the ending selector, but not including it" (section 4.2.8).

Refinement chains selectors, "where each refines the results of the previous one" (section 4.2.9). The specification's example is a TextQuoteSelector refining a FragmentSelector that identifies a paragraph. For evidence, the common pattern is a PDF page fragment refined by a text quote.

A note on browser text fragments

Browsers implement a different mechanism with a similar purpose: URL text directives of the form `#:~:text=[prefix-,]start[,end][,-suffix]`. The governing document, *URL Fragment Text Directives*, is a Draft Community Group Report (13 December 2023) that "is not a W3C Standard nor is it on the W3C Standards Track" (WICG). Text directives are convenient for a human reader, because a link scrolls to and highlights the passage. They are not a substitute for an annotation record: they carry no version, no state and no hash, and they are interpreted only by the browser at navigation time.

How each selector survives document change

The table summarises behaviour under four kinds of change, for a selector created against version 1 of a document.

Change between versions	TextQuoteSelector	TextPositionSelector	FragmentSelector (page=)	Text directive
Text inserted before the passage	Still matches	Points at wrong text, silently	Still matches if page unchanged	Still matches
Passage itself edited	Fails to match (detectable)	Points at edited text, silently	Page found; passage may differ	Fails to match
Same sentence repeated elsewhere	Ambiguous unless prefix/suffix disambiguate	Unaffected	Unaffected	Ambiguous unless prefix/suffix given
Re-pagination or reformatting	Matches after normalisation	Offsets may change	Page may change	Usually matches

The pattern is that quote-based anchors fail loudly and position-based anchors fail silently. For evidence, a loud failure is preferable: it tells the reviewer that the support needs to be re-examined.

Retained copies, states and hashes

Selectors find a passage in *a* representation. The evidence question is whether it is the representation that was read. The model answers this with State. A `TimeState` "records the time at which the resource is appropriate for the Annotation", with `sourceDate` (an `xsd:dateTime` that "MUST use the UTC timezone expressed as 'Z'") and `cached`, "a link to a copy of the Source resource's representation" (section 4.3.1). The specification notes that the timestamp "could be resolved via the Memento protocol", defined in RFC 7089, which provides datetime negotiation (`Accept-Datetime`, `Memento-Datetime`) between an original resource and its archived states.

Two hashes complete the record:

- **A hash of the retained representation**, so that a reviewer can confirm the cached copy is the one that was read.
- **A hash of the annotation itself**, so that the claim-to-passage link cannot be edited unnoticed. JSON does not have a single byte form; RFC 8785, the JSON Canonicalization Scheme, exists because "cryptographic operations like hashing and signing need the data to be expressed in an invariant format". It is an Informational RFC; canonicalising with it before hashing makes the hash reproducible across systems.

A worked example

The source is a fictional document, the *Annual Quality Report 2025* of Halden Therapeutics AG (a fictional company). After normalisation, version 1 reads:

> Halden Therapeutics AG Annual Quality Report 2025. Section 3. Batch release. In 2025 the Basel site released 412 batches. Two batches were rejected after release testing. Section 4. Deviations.

The claim to be supported is "Basel released 412 batches in 2025". The evidence span is the sentence "In 2025 the Basel site released 412 batches.", which occupies code points 77 to 121 of the normalised text. The annotation combines a quote selector, a position selector and a time state:

```
{
  "@context": "http://www.w3.org/ns/anno.jsonld",
  "id": "urn:example:anno:claim-0042",
  "type": "Annotation",
  "motivation": "assessing",
  "body": {"type": "TextualBody", "value": "Basel released 412 batches in 2025"},
  "target": {
    "source": "https://example.org/halden/quality-report-2025",
    "state": {
      "type": "TimeState",
      "sourceDate": "2026-02-10T08:30:00Z",
```

```
      "cached": "https://archive.example.org/halden/quality-report-2025/2026-02-10"
    },
    "selector": [
      {"type": "TextQuoteSelector",
       "prefix": " Batch release. ",
       "exact": "In 2025 the Basel site released 412 batches.",
       "suffix": " Two batches were"},
      {"type": "TextPositionSelector", "start": 77, "end": 121}
    ]
  }
}
```

The organisation also stores, outside the annotation, a SHA-256 of the cached representation and of the RFC 8785 canonical form of the annotation.

Now the report is revised twice.

Version	Change	Quote selector	Position selector	Correct conclusion
v2	Title gains " (revised)"	Matches at offset 87	Offsets 77-121 now return " release. In 2025 the Basel site released 41"	Support unchanged; update the position
v3	Figure corrected to 409	No match	Returns shifted text, no error	Support withdrawn; claim must be reviewed

Against version 2, a system relying on positions alone would extract a fragment that no longer forms the sentence and might still report a "match" if it only checked that the range was non-empty. The quote selector finds the sentence ten characters later; the system can record the new offset and keep the claim. Against version 3, the quote selector fails, which is the correct signal: the number changed. The cached copy dated 10 February 2026 still shows what was read, so the history of the claim remains defensible, while its current support does not. The site's article on preserving evidence when a cited source changes sets out the dispositions for each of these states.

PDF and HTML differences

HTML. The DOM is not the text. Before computing quotes or offsets, apply the normalisation the model requires (tags removed, entities decoded) and record which extraction was used, because two HTML-to-text converters can produce different whitespace and therefore different offsets. Dynamic content is a hazard the specification names explicitly ("dynamically transcluded content may change the selection").

PDF. Text extraction order may not follow reading order, hyphenation may split words across lines, and ligatures may be encoded as single glyphs. A robust pattern is a `FragmentSelector` with `page=` (per RFC 8118) refined by a `TextQuoteSelector` computed on that page's extracted text, plus the extractor name and version. For scanned PDFs, the quote is against OCR output and inherits its errors; record that.

Measuring whether citations support claims

Anchoring is a precondition for evaluating citation quality, not a measurement of it. Research benchmarks take the passage as given and ask whether it supports the statement. The Attributable to Identified Sources (AIS) framework of Rashkin et al. defines an evaluation for "whether model-generated statements are supported by underlying sources". The ALCE benchmark of Gao et al. evaluates citation quality automatically and reports that "on the ELI5 dataset, even the best models lack complete citation support 50% of the time". Both depend on knowing exactly which passage was cited; a selector-based record provides that. The site's method note on citing at the sentence describes the citation granularity this article encodes.

A verification routine

For each stored claim-to-passage annotation:

1. **Resolve state.** Retrieve the cached representation; verify its hash.
2. **Re-anchor historically.** Apply the quote selector to the cached text; confirm it matches exactly once and that the position selector agrees.
3. **Re-anchor currently.** Fetch the current representation and apply the quote selector.
4. **Classify.** Match at same offset: unchanged. Match at different offset: moved, update position. No match: changed, route for review. Multiple matches: ambiguous, extend prefix and suffix. Source unavailable: record as unavailable, not as unsupported.
5. **Record.** Write the outcome, date and new state as a new annotation version; do not overwrite the original.

Summary

A defensible citation to a passage is a small structured record: a source, a state that pins the representation and points to a retained copy, and at least two selectors, one quote-based to fail loudly when the passage changes and one position-based for efficient lookup and restricted content. Hashes of the copy and of the canonicalised annotation make the record checkable. The W3C Web Annotation model already defines all of it.

Sources

1. W3C, *Web Annotation Data Model*, W3C Recommendation, 23 February 2017. https://www.w3.org/TR/annotation-model/
2. Web Platform Incubator Community Group, *URL Fragment Text Directives*, Draft Community Group Report, 13 December 2023. https://wicg.github.io/scroll-to-text-fragment/
3. IETF, RFC 5147, *URI Fragment Identifiers for the text/plain Media Type*. https://www.rfc-editor.org/rfc/rfc5147
4. IETF, RFC 8118, *The application/pdf Media Type*. https://www.rfc-editor.org/rfc/rfc8118
5. IETF, RFC 7089, *HTTP Framework for Time-Based Access to Resource States — Memento*. https://www.rfc-editor.org/rfc/rfc7089
6. IETF, RFC 8785, *JSON Canonicalization Scheme (JCS)*. https://www.rfc-editor.org/rfc/rfc8785

7. Rashkin et al., *Measuring Attribution in Natural Language Generation Models*, arXiv:2112.12870. https://arxiv.org/abs/2112.12870
8. Gao et al., *Enabling Large Language Models to Generate Text with Citations*, EMNLP 2023, arXiv:2305.14627. https://arxiv.org/abs/2305.14627

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/annotation-model
2. https://www.rfc-editor.org/rfc/rfc5147
3. https://www.rfc-editor.org/rfc/rfc8118
4. https://wicg.github.io/scroll-to-text-fragment
5. https://www.rfc-editor.org/rfc/rfc7089
6. https://www.rfc-editor.org/rfc/rfc8785
7. https://arxiv.org/abs/2112.12870
8. https://arxiv.org/abs/2305.14627

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

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.