Embedded Metadata in “Plain” Text

Peter Constable, Joshua Hadley
January 13, 2026

Status: *This document provides background information for a topic we would like to have discussed by UTC. No specific proposals or recommendations are made.*

It was recently brought to our attention that the C2PA Content Credential Technical Specification has included a scheme for embedding provenance manifest metadata into unstructured (i.e., plain) text data using Unicode variation selector characters (see Appendix A.7). The scheme involves a simple transcoding of bytes for binary metadata using code points for the 256 variation selector characters. Variation selectors were explicitly chosen “because they are specifically designed to be visually non-rendering while remaining part of the valid Unicode character set”. The design aims to use Unicode characters so that “plain” text can still be processed as plain text (e.g., in copy/paste operations) while also providing a means of transporting the provenance metadata.

Besides the rise of generative AI and general societal concerns, a specific factor that could be driving urgency to create such a scheme is the EU Artificial Intelligence Act, which goes into force in August 2026. Article 50 states legal obligations for generated content, and penalties for non-compliance are high.[1]

> “Providers of AI systems, including general-purpose AI systems, generating synthetic audio, image, video or text content, shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated. …
>
> “Deployers of an AI system that generates or manipulates text which is published with the purpose of informing the public on matters of public interest shall disclose that the text has been artificially generated or manipulated. …”

For those of us who work on Unicode, we might interpret these statements, in relation to text content, as necessarily implying *structured* content formats (PDF, HTML, etc.) rather than plain text. EU legislators and regulators might not share our technical assumptions, however, and might expect that even a .txt file should include obligatory metadata.

The C2PA scheme uses valid Unicode characters, but in a non-conformant way: D1 in section 3.3.2 of the Unicode Standard stipulates that specifications for variation selectors in 23.4 are normative requirements, and the C2PA scheme violates the requirements for variation sequences in 23.4. (Note, though, that the discussion in 23.4 does not explicitly mention conformance.)

[1] €35M, or 3% of global annual turnover, whichever is higher. See Article 99.

At UTC 185, we briefly reviewed a proposal (L2/25-241) for two control characters to be used in plain text to make AI-related declarations. UTC generally concurred with the comments in the SEW report (section 5.5):

> “The SEW considers this information to be a higher-level markup. It is noted that control and format characters are generally not a good solution for similar purposes, as they may break text processing (search, text shaping, etc.). Unicode has deprecated or abandoned several such attempts, including language tags (U+E0001) or interlinear annotations (U+FFF9..U+FFFB).”

The outcome of UTC discussion was this action item:

> **[185-A116] Action Item for** Deborah Anderson, PAG: Draft an FAQ explaining the problems of mechanisms for interchanging metadata by means of invisible control characters.

If the C2PA working group had approached Unicode for feedback on the design of their scheme, we might have pointed to the above comments, but that likely would not have been an adequate response: their design recommends embedding the metadata in a way that is unlikely to adversely affect search or text shaping.

In Article 50 of the EU AI Act, Clause 7 refers to the EU AI Office “drawing up codes of practice” for how the law will get implemented. The AI Office has set up meetings for this purpose with several companies / agencies participating, including companies that are full-member participants in UTC. In that context, Microsoft, at least, will be arguing that plain text should be out of scope for enforcement of Article 50.

All of this raises some questions:

- Is the core spec clear enough in statements about conformance?
- Will an FAQ item resulting from 185-A116 provide adequate explanation of why attempting to embed metadata within plain text not a good idea?
- Should Unicode be engaging with the EU AI Office on the Code of Practice?
- Should we seek more interaction between Unicode and C2PA (perhaps a liaison relationship)?