Step Two: The Source-Document Catalog

Stated by Walter Wilkinson, August 2026 · companion to The Dispersal Hypothesis · Cryptographic Audit series

The proposal (as presented to CUT)

Step one is the digitization itself — the scan of the ~1,000-ft canvas. Step two, the first **use** of that scan: identify every text chunk affixed to the canvas. For each fragment, name the work it came from, with copious metadata: title, author or pen-name, edition, year, page and line location in the source. AI identifies text fragments from any century in a moment — so this step is fast, cheap, and can be demonstrated to CUT in advance on known passages.

The working assumption

The canvas carries the **complete** inventory of source documents Owen used to compose his five volumes. The catalog therefore delimits, for the first time, the actual corpus of the Word Cipher — including which works appeared under Bacon's own name and which under the bevy of pen-names.

What the catalog feeds

Downstream use	How the catalog serves it
The Dispersal Hypothesis test	Defines the empirical corpus (smooth source vs. seamed hosts) — replaces Owen's published citations with the artifact's own testimony
The "~40 works" claim	Cross-checked with a hard count, for the first time
The audit dataset itself	Combined with Owen's ~6-color pencil coding, the fragment table (work · edition · page · line · inch-coordinate · color) IS the dataset — per the project's master strategy
The masks cross-validation	The catalog is the independent list the Barsi-Greene masks prediction is graded against

Continuity note

> This sharpens an insight already in the project log from an earlier session: "OCR only enough to catalog which famous books the text chunks come from — texts like the First Folio are already perfectly known." Identification needs a handful of legible words per fragment, not full transcription.