Reading a journal that has no text: layout-aware transcription of scanned periodicals

Revised with a review of the literature and a measured audit

Internet History Initiative

July 2026 (revision 2)

The problem

ConneXions — The Interoperability Report (1987–1996) came to the collection as page scans with no text layer at all: not poor OCR, but none. One hundred fifteen issues, 3,573 pages, unreadable to search and to any text-side apparatus. The pages are 1990s desktop publishing: a wide margin column carrying side-headings, a main body column, boxed insets, mastheads, rules, and figures. Whole-page transcription by a vision model handles such pages poorly — reading order scrambles across the columns, and figure axis labels leak into the prose as token soup.

This note describes what worked: infer the page's rectangles first, transcribe each region separately at crop resolution, and reassemble in reading order with the region structure tagged. A vocabulary of the journal's own era, drawn from the collection, is supplied to the model as a prior. The whole corpus cost about two hours on two mid-size vision-language model instances (Qwen 3.6 35B), roughly 65 seconds per page single-threaded. This revision adds what the first version lacked: the historical context of the material, a review of forty years of prior art we initially built without, and a measured audit of our own output.

Historical context: preserving the desktop-publishing moment

There is a pleasing symmetry in this corpus. ConneXions was edited by Ole Jacobsen and published alongside the Interop conferences, documenting the protocols of the young Internet — and it was *produced* with the desktop-publishing tools of its own moment: the PageMaker-and-PostScript era that put professional multi-column layout, rules, pull-quotes, and margin apparatus within reach of a small trade publisher. The very features that made the journal handsome in 1990 — the wide scholia margin with side-headings, the boxed insets, the decorative rules — are the features that defeated optical character recognition for the following three decades. In preserving ConneXions we are preserving a layout idiom, not just a text; and the idiom is the hard part.

The large digitization programs met this class of material at scale and made strategic peace with it in different ways. The U.S. National Digital Newspaper Program (Chronicling America) chose page-level access only — article segmentation roughly doubles per-page cost — and accepted OCR whose quality "varies so widely it is nearly impossible to summarize"; the debt is being repaid today by machine-learning retrofits over twenty million pages. Australia's Trove paid

human contractors to zone every article and rekey every headline to 99% accuracy, then let the public correct the body text — the only program that solved "continued on page N" at scale, and it did so with people. The Europeana Newspapers project measured what happens when you skip the layout problem: raw character accuracy was tolerable, but *strict serialized* word accuracy on complex pages was about 20%, because reading-order and segmentation errors destroy the serialization even when the letters are right. That number is the whole argument for layout-first pipelines in one statistic.

Worked example

Figure 1 is an untouched page image; Figure 2 shows the inferred rectangles. Blue boxes are margin-column regions, red the body column, green full-width bands. The segmentation is classical image processing, not machine learning: ink profiles locate the column gutter (computed after excluding full-width decorative rules, which otherwise hide it); the gutter is cut wider than the morphological closing kernel so the columns cannot re-merge; scan borders are masked (they otherwise connect everything into one region); blocks then merge vertically only when they overlap horizontally. Each of those clauses earned its place by failing first — our first pass produced 68 fragments, our second one giant region containing the entire page.

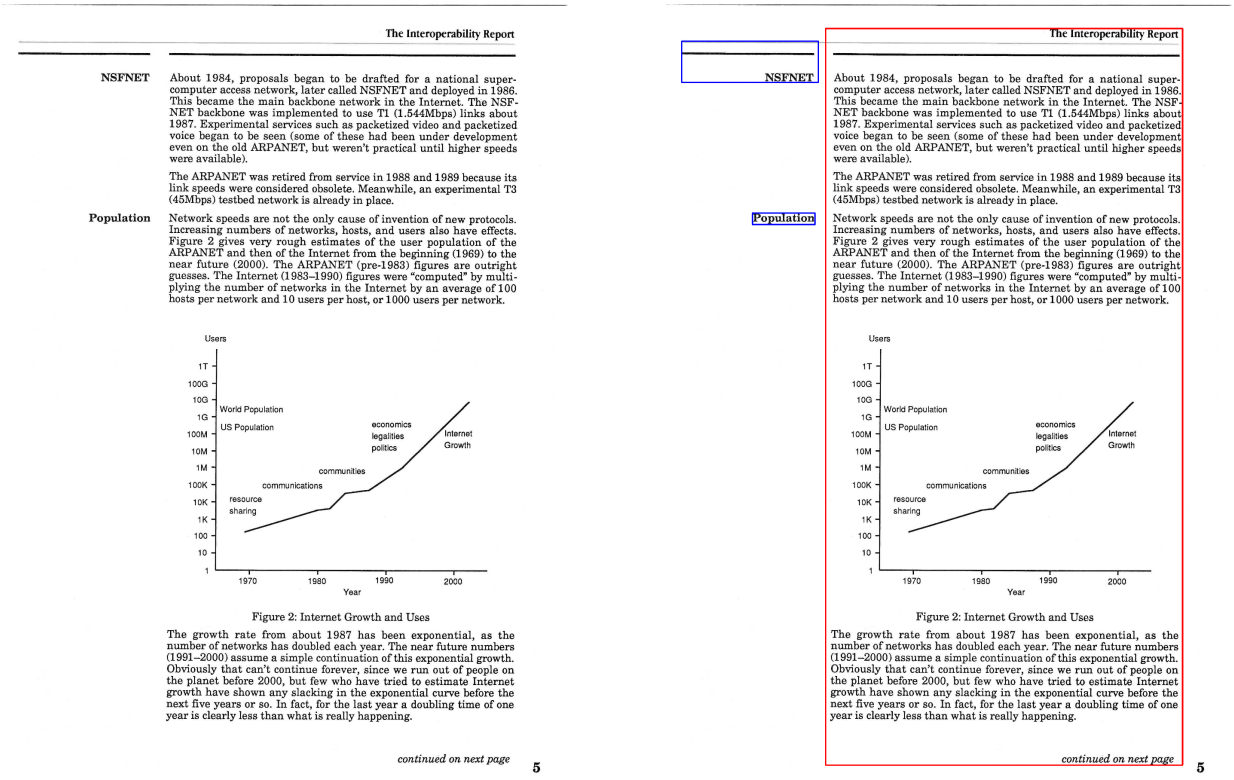

Figure 1: A ConneXions page (Vol. 5 No. 7, 1991) and its inferred regions.

We later learned that every one of those hard-won clauses reproduces a finding in the classical literature (below). We would have arrived faster by reading first; we understand the clauses better for having bled for them. Both things are true.

What the vocabulary and figure-awareness change

We transcribed the same body-column crop twice: once with a neutral "transcribe this faithfully" prompt, once with the production prompt, which adds (a) the journal's era vocabulary — protocol names, organization names, and personal names that the collection holds in quantity (OSI, CLNP, CMOT, SMDS, NSFNET, MILNET; Postel, Cerf, Jacobsen, …) — with the instruction to prefer these readings over similar-looking modern words but never to invent text; and (b) an instruction to describe figures rather than transcribe their fragments.

The differences on this one region, verbatim:

Neutral prompt — the figure arrives as axis-label soup interleaved with the prose, and a full paragraph of body text is dropped:

```
...legalities
politics
Internet
Growth
1970
1980
1990
2000
Year
Figure 2: Internet Growth and Uses
```

Production prompt — the dropped paragraph is present ("The ARPANET was retired from service…an experimental T3 (45Mbps) testbed network is already in place.") and the figure becomes a description a reader and a search engine can use:

```
FIGURE: A line graph titled "Figure 2: Internet Growth and Uses"
showing the growth of users from 1970 to 2000. The Y-axis is
logarithmic, labeled "Users" ranging from 1 to 1T. ... Reference
lines for "World Population" (near 1G) and "US Population" (near
100M) are shown. Annotations along the curve indicate phases of
use: "resource sharing" (1970s), "communications" (late 1970s),
"communities" (mid-1980s), and "economics, legalities, politics"
(late 1980s/early 1990s).
```

On this region the production prompt recovered 22% more body text and converted the figure from noise into signal. The era-vocabulary idea has a strong precedent we did not know when we used it: OverProof (Evershed & Fitch, 2014) corrected historical newspaper OCR with period-filtered word and n-gram language models and measured a 69–73% word-error reduction at corpus scale — context beat character-confusion modeling, hence their title. The same literature carries the caution: in the ICDAR post-OCR correction competitions only about half the submitted methods improved the text at all, and a 2025 study found LLM "repair" can *silently modernize* period vocabulary — fluency is a liability when the goal is fidelity. Our rule — vocabulary as prior, never substitution; spot-audit against images — is the right shape, but post-correction must always be gated on measured improvement.

What the literature would have told us

After the burn we commissioned a proper survey of the prior art — classical page-segmentation algorithms, the digitization programs, and the 2024–26 generation of document AI. Three find-

ings matter most.

The architecture is convergent. Decoupled layout-then-recognition is what the current benchmark leaders do: MinerU2.5 runs a fast global layout pass on a downsampled page, then feeds native-resolution crops of each region back through the same model — which is our pipeline with a learned detector in place of our ink profiles. Pure whole-page end-to-end decoding is effectively dead for production: Nougat's paper documents unrecoverable repetition degeneration on 1.5% of even in-domain pages, and every serious system since carries anti-repetition machinery (olmOCR detects a missing end-of-sequence and retries up a temperature ladder). Our classical segmenter is the component the field has since out-engineered: a small learned layout detector is the drop-in upgrade, and the crop-to-model stage keeps its value unchanged.

Every clause we bled for has a name. Scan-border masking: projection-profile methods degenerate to one whole-page region under border noise (the documented failure of XY-cut). Rule exclusion before gutter-finding: run-length smearing merges text through decorative bars. Gutter validation: Breuel's whitespace-cover column finder requires character-adjacency on both sides of a candidate gutter for the same reason our gutter needed to out-cut the closing kernel. Our giant-region failure and our 68-fragment failure are the two canonical failure directions (merge and split) of the PRImA evaluation taxonomy. Ray Smith's tab-stop paper (the Tesseract layout engine) is a compact catalogue of exactly our page family — cross-column headings, pull-outs, changing column counts — with a five-rule typed reading-order system worth stealing outright. And eynollah, built for the Berlin State Library, treats *marginalia* and *drop capitals* as first-class region classes — the only open system that does, and independent confirmation that our margin apparatus deserved its own tags.

Measure with reading-order-tolerant metrics, and keep the furniture. At the RDCL2017 competition Tesseract scored 48% raw character accuracy but 94% when the metric tolerated reading-order permutation — the text was right and the *order* was wrong, which plain edit distance cannot distinguish. Any evaluation of this class of work needs both numbers (the PRImA "Flex Character Accuracy"). And on the current benchmarks, headers-and-footers is the *worst-scoring category industry-wide*: mainstream systems either delete page furniture by design (olmOCR) or set it aside untranscribed. Our `[MARGIN:]`/`[BAND:]` tags — preserve and type, never delete — are genuinely differentiated, and they have a distinguished ancestor: the DAISY talking-book standard formalized *skippable* structures (sidebars, footnotes, page numbers a listener can toggle) decades ago. Our tags are that idea, rediscovered; the described editions now under construction return them to their original purpose.

Measured against ourselves: a 791-page audit

Sixteen reviewing agents compared 791 sampled pages (every issue of ConneXions and of the born-digital Internet Protocol Journal) image-against-stored-extraction, grading archival fidelity and a listener's-experience question — could a blind patron reconstruct the page's structure from our record alone? On ConneXions, 95.8% of 424 sampled pages graded A or B on fidelity. The defect classes found, in order of listener harm: margin side-headings captured but re-ordered to the end of the text instead of interleaved beside their paragraphs (the top complaint); occasional wholesale duplication of masthead/contents blocks; monospace code listings shredded into word salad; one subscription coupon dropped entirely; and roughly 8% of pages with hallucination-suspect passages — the class blind readers are least equipped to detect, and

therefore our first priority for the grounding-based confidence checks the literature recommends (verify model output tokens against a conventional-OCR lexicon of the same page; aggregate log-probabilities where the serving stack exposes them).

The audit also caught a pipeline-scale defect invisible at spot-check scale: a 900-token generation ceiling had truncated about 7,100 dense pages mid-transcription — *before* their figure description was emitted, silently deleting exactly the content this whole method exists to produce. All flagged pages (17,597 including the other classes) were re-burned at a higher ceiling the same day. Two standing rules came out of the incident: the renderer never clips a figure description (the transcription clips instead), and the ingest layer never lets a row that carries a figure description be replaced by one that lacks it.

Lessons learned

1. **Read the literature first; keep the bleeding optional.** Every segmentation clause we derived empirically was documented decades ago, with sharper tooling attached.
2. **Layout first is right, and now standard.** Keep crop-to-model; swap the hand-tuned segmenter for a learned detector when the next layout family arrives.
3. **Serialized edit distance lies about layout work.** Score with order-tolerant metrics alongside strict ones; otherwise you punish recognition for segmentation's sins.
4. **Token budgets are preservation policy.** A generation ceiling is a silent censor of whatever comes last in the output; put the most valuable content structurally first, or cap generously and audit for missing end-of-sequence.
5. **Page furniture is content.** Preserve and type it; the skippable/escapable semantics of DAISY are the publication-grade way to serve it.
6. **Fluency is not fidelity.** Era vocabulary as prior, never substitution; gate any model-based repair on measured improvement; build hallucination checks a blind reader could trust.
7. **Audit by looking.** A few hundred image-versus-extraction comparisons found defect classes (and one fleet-scale truncation bug) that no aggregate statistic surfaced.

Complete page result

The appendix below is the full stored transcription of a different page — "The RFC Story" from the first issue (1987) — exactly as it entered the collection's stores: margin regions tagged, body in reading order, figures described. Every claim in this note can be audited against the page images in the reading room, where each issue is readable page-by-page beside its transcription, and where each page now offers a "describe this page" disclosure: description first, verbatim text second, provenance voiced — the machine's words are marked as the librarian's, never the author's.

Honest limits

Region inference is tuned to this journal's layout family; a different publication needs its parameters re-verified against real pages (our verification loop — render, overlay the rectangles, look — takes minutes and is the step we would not skip). Tables transcribe as prose or description, not structure (TableFormer-class models exist and are the known upgrade). The model occasionally normalizes 1980s typography (drop caps arrive attached to the wrong word). Margin interleaving — placing each side-heading beside its paragraph rather than after the body — is designed but not yet shipped. Hallucination detection is currently spot-audit plus the grounding checks described above, not yet systematic. And the era vocabulary is a prior, not a leash: the two-column A/B above remains our only controlled prompt measurement; the 791-page audit is our only output-quality measurement. Both should grow.

Sources

The load-bearing prior art, for the reading list: Shafait, Keysers & Breuel, *Performance Comparison of Six Algorithms for Page Segmentation* (DAS 2006); R. Smith, *Hybrid Page Layout Analysis via Tab-Stop Detection* (ICDAR 2009); Kise et al., area-Voronoi segmentation (CVIU 1998); Antonacopoulos et al., the PRImA/ICDAR layout competitions and evaluation profiles (RDCL 2017/2019); Clausner et al., Flex Character Accuracy; eynollah (Berlin State Library); dhSegment (EPFL); Europeana Newspapers evaluation (HIP 2015); Holley, *How Good Can It Get?* (D-Lib 2009) and the Trove correction program; Evershed & Fitch, OverProof (DATeCH 2014); ICDAR post-OCR correction competitions (2017, 2019); olmOCR and olmOCR 2 (Allen AI); MinerU2.5; Nougat (Meta); Docling (IBM); OmniDocBench; the DAISY/NISO Z39.86 skippable-structure specifications; DIAGRAM Center image description guidelines; Lundgard & Satyanarayan on chart description levels (2021); WebAIM screen-reader surveys.

Appendix: a complete extracted page

```
[MARGIN] FIGURE: A horizontal line spanning the width of the page, with a thicker, textured
↪  horizontal line below it.

The Interoperability Report

The RFC Story

If you want to read the gospel on a particular protocol, you're likely to be pointed to an RFC.
↪  Before a proposed protocol is accepted for use in the Internet, it is discussed, reviewed
↪  and often revised by members of the Internet Activities Board (see separate story) and other
↪  interested parties. This dialogue is captured in a set of technical notes called Requests
↪  For Comments (RFCs). Their keeper and publisher is Dr. Jon Postel of USC-ISI.

A reading of all the RFCs paints a fascinating picture of the development of the Internet. RFC
↪  001, dated April 7th, 1969, was written by Steve Crocker and is entitled "Host Software".
↪  The Arpanet was just beginning to become a reality. BBN had already specified the software
↪  of the IMPs (nowadays called PSNs), and it was now time for the HOST working group to define
↪  the network software that would run on the hosts. To quote from RFC 001:

"During the summer of 1968, representatives from the initial four sites met several times to
↪  discuss HOST software and initial experiments on the network. There emerged from these
↪  meetings a working group of three, Steve Carr from Utah, Jeff Rulifson from SRI, and Steve
↪  Crocker of UCLA, who met during the fall and winter. The most recent meeting was in the last
↪  week of March in Utah. Also present was Bill Duvall of SRI who has recently started working
↪  with Jeff Rulifson.
```

Somewhat independently, Gerard De Loche of UCLA has been working on the HOST-IMP interface.

I present here some of the tentative agreements reached and some of the open questions
↪ encountered. Very little of what is here is firm and reactions are expected."

Many of the issued RFCs have become accepted standards without further ado, a phenomenon which
↪ has led some cynics to say that RFC stands for "Requirements For Compliance". The authors of
↪ RFCs have from time to time displayed some sense of humor. One that comes to mind is Mark
↪ Crispin's "Telnet-Randomly-Lose Option" (RFC 748) issued on April 1, 1978. But for the most
↪ part, RFCs contain technical specifications, usually written such that programmers can
↪ actually implement protocols from them.

In some cases the RFC will become an "Official ARPA Internet Protocol" and will be added to a
↪ list of such. The official list is itself an RFC which is issued periodically. (Currently
↪ RFC 991). This RFC also contains important comments about the current set of protocols and
↪ should definitely be read in conjunction with other RFCs.

[MARGIN] RFC 001 "Host
Software"

[MARGIN] Technical
specifications
aimed at
implementors

continued next page