PressMint QuickCheck: Operationalising Readiness Diagnostics for Interoperable Historical Newspaper Corpora

Elena Battaner Moro[1], Almudena Caballos Villar[2], María Cuevas Riaño[2], Marina Miguez Lamanuzzi[2], Dolores Romero López[2]

[1]Universidad Rey Juan Carlos (Madrid, Spain), [2]Universidad Complutense de Madrid (Spain)
elena.battaner@urjc.es, a.caballo@ucm.es, mmcuevas@ucm.es, marimigu@ucm.es, dromerol@ucm.es

Abstract

PressMint QuickCheck is a lightweight, reproducible readiness diagnostic for historical newspaper collections. Given a candidate dataset (ZIP export or IIIF manifests), it detects which components are present, identifies interoperability-critical metadata gaps, and applies lightweight OCR sanity checks. It produces three standardised artefacts: a human-readable readiness report, a minimal normalised manifest (CSV), and a tentative v1 scorecard (suitability_score 0-4) for prioritisation across collections. The workflow is delivered as a Colab-first notebook (no installation required). A key design decision treats content_language and metadata_language declarations as first-class interoperability signals, reflecting the multilingual scope of PressMint and ParlaMint corpora projects

Keywords: historical newspapers; interoperability; readiness diagnostics; metadata quality; OCR triage; PressMint

1. Introduction

Historical newspapers support research on language change, political history, cultural circulation and everyday life. However, digitisation programmes have often prioritised access over computational interoperability, resulting in heterogeneous "library exports": PDF issues with embedded OCR, OCR text directories, ALTO/METS-ALTO packages, partial TEI encodings, or IIIF manifests with descriptive metadata and image pointers.

National digitisation programmes illustrate this variety: some portals distribute PDF+OCR exports with embedded full text, others use METS/ALTO packages with Dublin Core metadata, and some university heritage repositories expose IIIF Presentation API 3.0 manifests. Such variability complicates a basic but essential first step: deciding whether a collection is ready for interoperability-oriented processing and what remediation should occur before investing in conversion and annotation.

PressMint seeks to compile interoperable corpora of historical newspapers and enable comparable processing across collections, following the model of projects such as impresso (Ehrmann et al., 2020). Onboarding candidate collections remains resource-intensive, especially for small DH teams and libraries without dedicated engineering capacity. QuickCheck addresses this gap by providing a low-barrier readiness diagnostic that makes early assessment transparent, reproducible, and actionable.

2. Problem Statement and Contribution

We target an onboarding problem prior to model choice or annotation schemes: given a candidate collection, (1) what components are present (OCR, images, structured metadata), (2) which interoperability-critical metadata are missing or inconsistent (e.g., dates, stable identifiers, provenance, rights, language declarations), and (3) whether OCR exhibits obvious risk signals (empty or low-signal items) that would undermine downstream NLP. OCR quality in historical collections is known to vary considerably (Springmann and Lüdeling, 2017). In many projects this information is produced ad hoc and cannot be compared across collections.

Our contribution is threefold:

- A staged readiness checklist that operationalises these concerns into concrete checks, with explicit treatment of content_language and metadata_language declarations as first-class interoperability signals.
- A lightweight reference implementation (Colab-first notebook) that produces three standardised outputs: a readiness report, a minimal normalised manifest (CSV), and a tentative v1 scorecard for prioritisation.
- A governance-aligned schema and scoring design intended as a starting point for community co-design within the PressMint ecosystem, with configurable weights and thresholds subject to community review.

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 11–15

3. PressMint QuickCheck Method

3.1 Workflow overview

The workflow ingests a single ZIP package or a folder of IIIF manifest JSON files, detects which components are present, applies staged checks, and writes three artefacts: a readiness report (HTML), a minimal manifest (CSV), and a scorecard (JSON with optional per-check breakdown CSV). The pipeline is issue-level: each row in the manifest corresponds to one newspaper issue, defined as a single PDF file, a folder of OCR/ALTO files, or a unit identified heuristically from filenames.

3.2 Inputs and graceful degradation

QuickCheck detects which components are present and applies only the corresponding checks. This design supports realistic library deliveries: collections may include PDF+OCR but no structured metadata, or IIIF manifests with rich descriptive fields but no OCR export. When inputs are missing or partial, QuickCheck still produces an inventory and a minimal manifest while explicitly reporting gaps via structured flags.

3.3 Readiness checklist

Checks are grouped into four lightweight families: Inventory and structure: file tree, sizes, detected components, and missing-component flags.

- Metadata completeness and consistency: field coverage, parseable dates (ISO 8601), stable identifiers, content_language and metadata_language declarations (presence check, basic BCP 47-like syntax validation; missing declarations flagged as a high-impact P1 readiness gap), duplicates, and other high-impact missing fields.
- OCR sanity (if present): coverage, empty or low-signal items, abnormal character ratios, and extreme-length outliers (triage indicators only; not a measure of OCR accuracy).
- IIIF checks (if present): basic manifest validity, canvas counts, label presence, and consistency of identifiers and links, following the IIIF Presentation API 3.0 specification (IIIF Consortium, n.d.).

3.4 Prioritisation: needs_priority and needs_actions

Each issue receives a structured flag list (issues_flags) from which two fields are derived. needs_priority classifies urgency at issue level: P0 (blocker — the issue cannot be onboarded; triggered when no usable text or structure is present), P1 (required — onboarding possible after resolving specific gaps), or P2 (desirable — does not block onboarding). The final priority is the most severe flag present: a single P0 flag makes the issue P0 regardless of other signals.

needs_actions translates each flag into a concrete remediation step. Table 1 shows the flag-to-action mapping.

Table 1: Flag-to-action mapping (v2.0).

Flag	Priority	Action
P0:no_text_or_struct	P0	Provide OCR text (TXT/ALTO), IIIF manifests, or structured metadata exports
P1:missing_language_declaration	P1	Add declared content_language (and metadata_language) using BCP 47 tags
P1:unparseable_date	P1	Provide/normalise issue date (ISO 8601) or encode it in filenames/metadata
P1:missing_rights	P1	Add rights/licence statement at collection/issue level
P1:missing_provenance	P1	Add provider/source provenance label for cross-collection comparison
P1:ocr_empty_or_low_signal	P1	Check OCR extraction quality; consider re-OCR or alternative exports
P1:iiif_invalid	P1	Validate IIIF manifests (id/type/items/labels)
P2:generated_id	P2	Stable external ID

Flag	Priority	Action
		not found; generated from filename/date
P2:fields_from_defaults:X,Y	P2	Fields X, Y filled from notebook defaults, not declared by provider. Verify accuracy and request explicit metadata declarations.

The P2:fields_from_defaults flag (introduced in v2.0) addresses a transparency gap identified during external testing (see Section 4): when a user sets notebook-level defaults for content_language, rights, or source_provenance, these values populate the manifest but the provider has not declared them. In v1.9, this produced an empty needs_actions despite fields being synthetic — a misleading signal. In v2.0, the flag is added with the list of affected fields, ensuring needs_actions always reflects the actual state of provider-declared metadata.

4. Minimal manifest (CSV) for harmonisation

The manifest schema is intentionally minimal and designed as a harmonisation starting point for PressMint onboarding workflows. Language-related fields (content_language, metadata_language, language_present) are first-class outputs because PressMint and ParlaMint-style corpora are intrinsically multilingual and require explicit language signalling for interoperability. The suitability_score (0–100) and suitability_level (0–4) are derived from a weighted checklist; weights and thresholds are configurable and subject to community review aligned with PressMint governance decisions. The manifest schema is designed to align with PressMint operational requirements for onboarding: fields such as stable identifiers, rights statements, and language declarations correspond directly to minimum metadata requirements under discussion within the PressMint community.
Table 2 shows the minimal manifest schema. Fields marked “Always” are present in every output row; “If present” fields are populated when the source collection declares them.

Table 2: Minimal manifest schema (v1).

Field	Status	Notes
item_id	Always	Original stable ID if present; generated otherwise (P2:generated_id flag)
date	Always	Issue date parsed to ISO 8601; empty with flag if unparseable
content_language	Always	Declared language of OCR/full-text content; BCP 47-like validation; missing declaration flagged as P1
metadata_language	If present	Declared language of descriptive metadata fields; BCP 47-like validation where present
language_present	Always	Boolean: true if at least one language declaration found; false triggers P1 flag
title_or_masthead	If present	Title or masthead string from metadata
source_provenance	If present	Provider/source label; key for cross-collection comparison
rights	If present	Rights/licence information when declared
files_present	Always	Detected components (PDF, OCR, ALTO, IIIF…)
issues_flags	Always	Structured flags for missing fields, broken

Field	Status	Notes
		references, low-signal OCR, etc.
suitability_score	Always	0–100 integer; derived from weighted checklist. Weights are v1, tentative, configurable
suitability_level	Always	0–4 coarse category: 0=Not usable; 1=Text-only triage; 2=Candidate; 3=Ready for harmonisation; 4=PressMint-ready
needs_priority	Always	P0/P1/P2 derived from issues_flags.
needs_actions	Always	Short remediation steps per flag (see Table 1).

4.1 Limitations

QuickCheck v1 does not attempt full PressMint TEI conversion, article segmentation, heavy NLP enrichment, or automatic language identification. It is explicitly a pre-flight diagnostic for onboarding, not a pipeline component. The suitability scorecard measures onboarding readiness only; it does not measure OCR accuracy, linguistic quality, scholarly value, or corpus relevance. The language_guess field (optional, v1) is a lightweight triage hint based on stopword frequency and is not a substitute for declared BCP 47 tags. Issue boundaries are determined heuristically from filenames when explicit boundaries are absent, which may misgroup files in collections with non-standard naming conventions. Scoring weights and suitability thresholds are v1 and tentative; they are configurable and subject to community governance rather than fixed as final technical decisions.

When collections lack reliable metadata — a question raised during peer review — QuickCheck still produces an inventory and a minimal manifest while flagging all gaps explicitly (P0/P1/P2). Notebook-level defaults can supply provisional values for missing fields, but v2.0 explicitly flags these as P2:fields_from_defaults to avoid masking the underlying metadata gap.

5. Demo and Use Case

The demo illustrates the end-to-end workflow on two sample packages representative of common library delivery patterns: (i) a set of PDF issues with embedded OCR from a national digital newspaper portal (PDF+OCR sample); and (ii) a set of IIIF manifests from a university heritage repository exposing IIIF Presentation API 3.0 (IIIF sample). We show: component detection and inventory across both delivery patterns, metadata coverage and identifier/date/language consistency, OCR sanity indicators on the PDF sample, and export of a unified minimal manifest covering both collections.

The demo then shows how the manifest supports (a) selecting a “first subset” for harmonisation, (b) prioritising remediation tasks such as missing language tags, unstable identifiers, or absent provenance/rights fields, and (c) producing comparable collection-level summaries across the two input types.

6. Discussion: Positioning within PressMint

QuickCheck is intended to complement, not replace, PressMint’s conversion and annotation pipelines. By standardising an early assessment step, it makes onboarding decisions more transparent and comparable across partner collections, supporting the interoperability goals pursued by initiatives such as Europeana Newspapers (Neudecker and Antonacopoulos, 2016) and the CLARIN PressMint flagship project (CLARIN ERIC, n.d.).

The minimal manifest schema is designed to align with PressMint’s emerging operational requirements for onboarding. Fields such as stable identifiers, rights statements, and language declarations correspond to minimum metadata requirements under discussion within the PressMint community, and the suitability_level 4 threshold (“PressMint-ready”) is intended to reflect those requirements as they are formalised through governance. This alignment is intentionally configurable rather than fixed, since PressMint’s operational decisions are still evolving.

Language field coverage is one check among others in the readiness checklist; it is included because missing or inconsistent language tags are a recurring interoperability gap in multilingual corpora. This is particularly valuable for small teams and library collaborations where engineering capacity is limited, and format heterogeneity is the norm. A shared readiness report and minimal manifest also provide a concrete basis for discussion between content providers and technical teams, helping to align

expectations and define realistic remediation plans.

The schema, checklist, and scoring are presented as a pathway for community co-design within the PressMint ecosystem. We do not claim prior endorsement by PressMint or ParlaMint governing bodies.

7. Availability

The notebook (in English and Spanish, other languages and formats are foreseen) and sample outputs are available in the following public repository URL: https://github.com/ebattanermoro/PressMint-Quickcheck.

8. Conclusion

PressMint QuickCheck operationalises a lightweight readiness diagnostic for historical newspaper collections and provides a reproducible, low-barrier reference workflow. By producing three standardised artefacts (a readiness report, a minimal harmonisation manifest, and a tentative v1 scorecard for prioritisation) it supports faster and more transparent onboarding of candidate collections into the PressMint ecosystem. Language declarations are treated as first-class interoperability signals, reflecting the multilingual scope of PressMint and ParlaMint corpora. The v2.0 release addresses a transparency gap identified during external testing, ensuring that notebook-level defaults are always explicitly flagged rather than silently masking missing provider metadata. We would like to thank three anonymous reviewers and the PressMint workshop team at LREC 2026 for their useful and encouraging comments and suggestions.

9. Bibliographical References

CLARIN ERIC. (n.d.). PressMint. https://www.clarin.eu/pressmint (accessed 25 March 2026).

Ehrmann, M., Romanello, M., Clematide, S., Strobel, P. B., and Barman, R. (2020). Language Resources for Historical Newspapers: the Impresso Collection. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pp. 958-968. Marseille, France. European Language Resources Association (ELRA). https://aclanthology.org/2020.lrec-1.121.pdf (accessed 25 March 2026).

IIIF Consortium. (n.d.). IIIF Presentation API 3.0. https://iiif.io/api/presentation/3.0/ (accessed 25 March 2026).

Neudecker, C. and Antonacopoulos, A. (2016). Making Europe's Historical Newspapers Searchable. In Proceedings of the 12th International Conference on Document Analysis Systems (DAS 2016), pp. 405-410. Santorini, Greece. IEEE. https://doi.org/10.1109/DAS.2016.83 (accessed 25 March 2026).

Springmann, U. and Lüdeling, A. (2017). OCR of Historical Printings with an Application to Building Diachronic Corpora: A Case Study Using the RIDGES Herbal Corpus. Digital Humanities Quarterly, 11(2). https://dhq.digitalhumanities.org/vol/11/2/000288/000288.html (accessed 25 March 2026).

Statement on the use of AI: Claude Sonnet 4.6 and ChatGPT 5.2. were used for this work to review texts and code.