Subject Index Evaluation: A Source-Grounded, Candidate-Blind Method

John Camden

Publication Intelligence, LLC

2026-08-29

Research methodology

Canonical article
indexpdf.com/research/subject-index-evaluation-methodology

Abstract

This article presents a methodology for evaluating finished subject indexes; it does not describe the separate process by which IndexPDF or any other system generates an index. The method fixes the source, scope, page map, policy, and failure gates; builds a candidate-blind benchmark from the complete supplied source; preserves the finished candidate without editorial repair; and performs locator, missing-access, and whole-index audits. A versioned calculation profile derives six dimensions from frozen ledgers through explicit credits, denominators, component weights, consequence caps, uncertainty bounds, and decimal rounding, while publication gates remain separate. A first full application to the 425-page body text of William Doyle's The Oxford History of the French Revolution audited 5,338 atomic locator claims and 1,366 source-derived subjects, producing a canonical as-delivered score of 72.5/100 and seven failed readiness gates. The result is an inspectable measurement system for comparing human, AI, and hybrid indexes against the source rather than treating any candidate as ground truth.

Keywords: subject-index evaluation; back-of-book indexes; candidate-blind benchmarking; locator precision; benchmark recall; index quality; information retrieval; AI evaluation

Suggested citation. Camden, John. 2026. Subject Index Evaluation: A Source-Grounded, Candidate-Blind Method. Publication Intelligence, LLC. indexpdf.com/research/subject-index-evaluation-methodology.

Contents

1 Scope and research problem

A subject index is both a set of factual claims and a navigation system. Each locator asserts that a source page substantively treats the meaning expressed by a complete heading path; the index as a whole asserts that important treatments can be found through intelligible vocabulary, hierarchy, and cross-references. Evaluation must therefore ask two different questions: Are the access points supplied by the index supported? and Does the index supply access to what an intended reader is likely to need?

This distinction separates subject-index **evaluation** from subject-index **generation**. Index-PDF concerns the creation of candidate indexes. Benchmark construction may begin before a candidate is delivered; candidate-specific judgment begins only when a finished candidate exists. The same evaluation logic applies to professional, author-created, AI-generated, and hybrid indexes, although source and candidate extraction may require format-specific adapters. Candidate-generation prompts, models, and workflows are deliberately excluded from benchmark construction.

The usual shortcut is to treat an existing published index as a gold standard and compute overlap. That design has enabled repeatable research on automatic back-of-book indexing (Csomai and Mihalcea, 2006, 2008; Wu et al., 2013), but agreement with one indexer is not identical to correctness. Indexers may validly differ in vocabulary, granularity, exhaustivity, hierarchy, and audience assumptions; they can also share or repeat omissions. Inter-indexer consistency studies and reviews report wide variation and identify indexing depth, concept selection, policy, and vocabulary specificity as sources of disagreement (Hurwitz, 1969; Rolling, 1981; Markey, 1984; Reich and Biever, 1991). A published index is consequently valuable evidence and a useful candidate, but it is not self-validating ground truth.

The methodology addresses that problem by making the indexed source-not another index-the primary evidentiary authority. Its central contribution is an end-to-end protocol that combines:

1. a policy-bound benchmark derived from every in-scope page in benchmark contexts that never see candidate material, before candidate judgment;
2. separate measurement of asserted-access reliability and missing access;
3. atomic verification of complete-heading-path/page claims;
4. global evaluation of hierarchy, terminology, cross-references, density, and reader navigation;
5. explicit uncertainty, adjudication, provenance, and critical failure gates; and
6. content-addressed artifacts that permit resumption, audit, and like-for-like comparison.

The contribution is the operational integration of these established ideas into a full-book, auditable protocol-not the invention of precision, recall, expert review, or indexing standards.

2 Prior art and the evaluation gap

2.1 Standards specify qualities, not a validation experiment

ISO 999:1996 and ANSI/NISO Z39.4-2021 provide the normative foundation for index content, organization, vocabulary, specificity, locators, cross-references, and presentation. At the date of writing, the 1996 ISO edition remained the published standard while a third-edition final draft was under approval. Anderson's earlier NISO technical report supplies extensive guidance on index design, term relationships, and syntax (Anderson, 1997). Professional guidance makes these expectations usable through checklists and award criteria (American Society for Indexing, 2015, 2025, n.d.; Marshall, 2023a, 2023b).

These sources define what a good index should accomplish, but their published scopes do not by themselves provide a complete validation protocol covering benchmark construction, atomic

audit units, missing-access measurement, uncertainty adjudication, candidate leakage, and noncompensatory decisions. Heuristics such as subdividing long locator strings remain useful diagnostics, not universal empirical laws.

2.2 Structural metrics and reader studies answer narrower questions

Book-index research has measured index length, subheading form, authorship effects, structural richness, and user preferences (Gratch et al., 1978; Wittmann, 1990; Diodato and Gandt, 1991; Diodato, 1994). Bennion (1980) treated a book and its index as an information-retrieval system, while Jörgensen and Liddy (1996) and Abdullah and Gibb (2008) connected index features to search tasks. Johncocks (2008) argued for quantitative diagnostics while retaining intellectual judgment; Quinn (2015) reviewed evaluation criteria and the Australian and New Zealand Society of Indexers' experience applying them. Coe's literature review found comparatively little empirical evidence about actual book-index use; her subsequent six-participant study identified overview and specific-topic lookup as important behaviors (Coe, 2014, 2015).

These studies justify measurable reader tasks and structural diagnostics, but neither raw structure nor standards compliance directly establishes source correctness. Live-user testing provides a distinct external validation of retrieval outcomes and is not replaced by the method presented here.

2.3 Automatic-index evaluation commonly measures resemblance

Csomai and Mihalcea's testbed paired books with their published indexes and applied precision, recall, and F-measure to index-term selection (2006, 2008). Aït El Mekki and Nazarenko (2006) explicitly observed that back-of-book indexes lack a simple objective reference and evaluated a human-machine workflow using descriptor and relation measures. Wu et al. (2013) advanced the unit of analysis from keywords toward term-locator pairs, yet still treated published indexes as ground truth and left full hierarchy, synonym handling, and cross-references outside scope. Golub et al.'s broader framework for automatic subject indexing argues for more than a single gold-standard method by triangulating direct assessment, workflow evaluation, and downstream retrieval (2016).

Adjacent research reaches compatible conclusions. Thesaurus-aware consistency measures can recognize semantic agreement missed by exact term identity (Medelyan and Witten, 2006). Large-scale keyphrase evaluation shows that conclusions depend on dataset, experimental setup, and reference provenance (Gallina et al., 2020), while KPEval separates reference agreement from source faithfulness, diversity, and downstream utility (Wu et al., 2024). These advances remain narrower than subject-index evaluation because they do not test page locators, hierarchy, cross-references, or book-scale navigation.

Recent journal articles and professional white papers directly examine generative-AI indexing outputs (Izzard, 2024; Bartmess and Combs, 2025, 2026; American Society for Indexing AI Committee, 2026). Collectively, they assess completeness, navigability, and accuracy under professional criteria rather than term overlap alone. The present method differs by preconstructing a candidate-blind source benchmark, auditing all candidate-to-source and source-to-candidate units, freezing evidence provenance, and applying noncompensatory release gates.

The present method extends that logic from isolated terms to a completed index. It evaluates the semantic relation among heading path, source passage, locator, and reader task, then evaluates the resulting access architecture as a whole.

2.4 Current commercial benchmarks measure different constructs

Commercial work is informative; this review identified no independent replication. IndexerLabs' Oxford demonstration controls index size, manually reconciles exact and semantic heading matches, and compares locators within matched headings. Its reported 608 matches between 1,058 generated and 1,066 printed top-level entries quantify convergence toward one published reference index. The company's later 120-run pruning study and 20,000-locator verification report provide first-party evidence about topic retention under pruning and locator checking. The public reports do not document an external, complete audit of the defined in-scope source for candidate-only headings, source omissions, and whole-index navigation.

Indexia's Indexing Standards Benchmark converts rules from ISO 999, ANSI/NISO Z39.4, and The Chicago Manual of Style into rule-specific checks. That is a substantive evaluation methodology, but it primarily measures compliance with professional rules. Rule compliance and source-grounded retrieval validity are complementary constructs.

Evaluation family	Primary reference	What it measures well	What it cannot establish by itself
Published-index overlap	One human-created index	Reproduction of selected terms or locators	Whether the reference omitted access; whether candidate-only access is valid
Standards compliance	Codified professional rules	Formal, structural, and stylistic conformance	Complete source coverage or substantive support for every locator
Source-grounded evaluation	Frozen source benchmark plus source evidence	Supported access, missing access, semantic fidelity, and navigation	Actual reader performance without a user study

3 Evaluation model and design principles

3.1 Two linked representations

The candidate index is represented without correction as a set of heading nodes, complete paths, displayed locators, expanded atomic locator assignments, and cross-references. The basic locator claim is:

$$c = (h, p),$$

where h is the full heading path-not an isolated label-and p is one logical source-page label. A range is not one claim: it is expanded through the ordered page map and every page is judged separately.

The source benchmark is a graph of retrieval requirements rather than a model index. Each stable subject record contains its meaning, priority, source stance, acceptable access routes, evidence pages, relationships, and reader tasks. This representation can accept synonymous headings, inversions, double postings, cross-reference routes, and alternative hierarchies when they provide equivalent access. It therefore avoids requiring a candidate to reproduce the wording or architecture of a single reference index.

3.2 Fixed principles

Principle	Operational rule	Bias or failure controlled
Candidate blindness	Discover, review, and freeze source requirements before candidate judgment	Post hoc benchmark tailoring and circularity
Source grounding	Require substantive treatment, not mere string occurrence	Keyword false positives and attribution-only access
Bidirectionality	Audit candidate claims and source requirements separately	Precise-but-sparse and broad-but-noisy indexes
Candidate fidelity	Preserve malformed text, hierarchy, ranges, duplicates, and references	Evaluator repair of the object being measured
Atomicity	Judge one complete path on one page	Range inflation, inherited parent support, and ambiguous compounds
Whole-index review	Assess hierarchy, distribution, terminology, references, and mechanics globally	Locally valid entries that form a poor navigation system
Precommitment	Freeze scope, page map, policy, thresholds, gates, and allowed deviations before candidate judgment; bind a versioned calculation profile before scoring	Unrecorded post hoc changes to evidence or arithmetic
Explicit uncertainty	Retain unknown and uninspectable states outside forced binaries	False precision and denominator distortion
Separate gates	Apply critical readiness tests without changing the numeric score	Compensation of fatal defects by unrelated strengths

Candidate blindness is a procedural safeguard, not proof of statistical independence. The implemented protocol uses a fresh candidate-unseen context for complete benchmark review and omission search. That separation does not substitute for multiple human raters or an inter-rater reliability study.

If candidate material enters a discovery or benchmark-review context, blindness is recorded as compromised and those stages must be rerun in a fresh candidate-unseen context before candidate-blind claims are published.

4 Protocol

The public workflow specification is a 16-stage state machine, grouped below into eight methodological phases.

The semantic steps may be performed by qualified human reviewers, model-assisted reviewers, or a hybrid team. In the Oxford application, language-model agents performed source and candidate review under human orchestration and adjudication; deterministic programs controlled identities, routing, validation, and arithmetic. Runtime separation prevented the candidate from entering benchmark-construction contexts, but it cannot establish that a pretrained model had never encountered the book or its index.

Phase	Required operation	Frozen output or invariant
1. Register	Identify source bytes and edition, intended readership, indexable scope, audit mode, and publication profile	Source hash and run identity
2. Map	Map every one-based PDF leaf to its printed or logical label; assign nonoverlapping page ownership to intellectual chunks	Page map and chunk manifest
3. Predeclare	Instantiate the content policy, uncertainty rules, density calibration, gates, and allowed deviations; declare audit mode	Hashed evaluation policy
4. Discover	Inspect every in-scope source page with the candidate unseen; record subjects, meanings, stance, evidence, relations, exclusions, uncertainties, and reader tasks	Chapter discovery artifacts
5. Synthesize and review	Reconcile cross-chapter concepts; review every subject, relationship, and task in fresh context; perform an omission pass and adjudicate defects	Frozen source benchmark
6. Prepare candidate	Extract and normalize the delivered index in isolation, accounting for every line, record, path, locator, and reference without editorial repair	Candidate representation, inventory, and benchmark lock
7. Audit	Verify every atomic locator; test access to every scored subject, treatment unit, and required task; evaluate whole-index structure	Locator, missing-access, and structure ledgers
8. Decide and publish	Bind the calculation profile; derive item diagnostics and six dimensions from frozen ledgers; apply gates; project a display-only report and validate it against the calculation	Calculation artifact, result, validation receipt, readiness status, and public report

4.1 Page identity before semantic judgment

PDF leaf numbers, printed page labels, and local pages in derived chapter PDFs are distinct coordinates. The method preserves logical labels as strings and records their exact relationship rather than assuming one global arithmetic offset. This accommodates Roman numerals, prefixed labels, plates, duplicated labels, and irregular segments. Chapter packets may include context pages, but each source page has exactly one judgment owner. Ranges are expanded by walking the frozen ordered map; reversed, cross-segment, or ambiguous ranges remain unresolved instead of being silently repaired.

This apparently mechanical stage is methodologically important. A semantically correct judgment attached to the wrong coordinate is not reproducible evidence.

4.2 Candidate-blind benchmark construction

Reviewers inspect the complete in-scope source and record substantively treated, plausibly searchable subjects. A passing mention, attribution-only name, citation, or isolated example is excluded by default; a short passage may qualify when it performs decisive argumentative work. Density is not consulted when discovering subjects and cannot be used to prune or pad the benchmark.

Chapter discoveries are synthesized into stable whole-source concepts. A fresh review then covers every subject, relation, and reader task and conducts a separate omission search. Critical, major, and uncertain items are adjudicated while original and revised judgments remain in the ledger. Only an approving full review freezes the benchmark. The resulting graph specifies acceptable retrieval outcomes, not a preferred finished index.

4.3 Candidate preparation without repair

Under the candidate-preparation protocol, preparation may run in an isolated mechanical context once source identity, edition, page map, chunks, policy, and audit mode are fixed; it may not consult benchmark subjects or make quality judgments. Its outputs are integrated only after benchmark freeze. Preparation preserves delivered spelling, accents, punctuation, hierarchy, duplicates, ranges, mixed locator/reference records, and unresolved structures. Provenance distinguishes six questions that are often conflated: byte identity, internal completeness, structural continuity, source-edition compatibility, locator-map compatibility, and fidelity to a separately authenticated copy.

This separation permits later sensitivity analysis. If a candidate was corrupted during transmission or reconstruction, the observed evaluation can remain immutable while a labeled counterfactual shows which conclusions would change under confirmed corrections.

4.4 Three complementary audits

Candidate to source. Every complete-heading-path/page assignment is classified as `supported`, `partially_supported`, `unsupported`, or `uninspectable`. Parent support does not imply child support. A compound heading is supported only when all components are substantively joined on the cited page.

Source to candidate. Every scored frozen subject, unique expected treatment unit, and required reader task is tested against the complete candidate. A treatment unit is the unique tuple of subject ID, document page, and locator class; duplicate evidence records are coalesced while retaining their evidence IDs. Access may be direct, indirect but usable, partial, or missing. Routine missing-access review uses the frozen benchmark, complete candidate, and canonical locator ledger; it does not reopen or silently reinterpret the source.

Whole index. A single global pass evaluates heading hierarchy, terminology, concept distribution, direct access, cross-references, fragmentation, locator density, mechanics, and navigation. This pass remains whole-index work because chapter-local judgments cannot establish global balance or reference topology.

One underlying defect is recorded once and linked to all affected measures. Uncertain judgments retain alternatives, confidence, and evidence needed; `uninspectable` is not recoded as failure or quietly removed. The default policy excludes uninspectable locators from binary precision denominators and discloses their rate, with a predeclared tolerance gate.

5 Measurement and decision rules

5.1 Supporting measures

The method publishes denominators and disaggregated measures rather than asking one score to carry every interpretation.

For inspectable locator claims, let s, p, and u denote supported, partially supported, and unsupported assignments. The method reports three distinct rates:

$$S_L = \frac{s}{s+p+u}, \qquad P_L = \frac{s}{s+u}, \qquad A_L = \frac{s+p}{s+p+u}.$$

These are the strict supported-locator rate, binary precision after setting partial cases aside, and at-least-partial support rate. Reporting all three prevents the treatment of partial cases from disappearing inside one number.

Benchmark concept access is priority weighted. Essential and major subjects receive weights 3 and 2. An optional subject receives weight 1 only when the frozen benchmark marks it as scored; otherwise it remains outside the denominator. Complete, partial, and missing access receive values 1, 0.5, and 0:

$$R_B = \frac{\sum_i w_i a_i}{\sum_i w_i}.$$

Unmeasured, unresolved, or uninspectable access is not silently assigned a zero; its treatment and the resulting denominator must be disclosed. The measure is more precisely described as benchmark concept coverage or discovered-access recall than as exhaustive recall of every possible valid subject. Expected-treatment recall uses unique treatment units rather than raw evidence records. If f and m are found and missed units, then

$$R_T = \frac{f}{f+m}.$$

The page-reference reliability dimension combines strict locator support S_L and expected-treatment recall R_T by the harmonic mean $F_R = 2S_L R_T/(S_L + R_T)$. This is valid because both terms concern page-level substantive retrieval, but the component rates remain public and F_R cannot override a critical gate. Strict and partial-credit reader-task success, cross-reference validity, conceptual and stance defects, clutter, and structural-defect counts likewise remain separately visible. The caution resembles information-retrieval evaluation with incomplete relevance judgments (Buckley and Voorhees, 2004). Graded-relevance metrics such as nDCG (Järvelin and Kekäläinen, 2002) are useful for future ranked retrieval experiments, but an alphabetic print index is not itself a relevance-ranked result list.

Density is calculated within each chapter or approved intellectual unit:

$$D_{P,c} = 1000\frac{P_c}{W_c}, \qquad D_{L,c} = 1000\frac{L_c}{W_c},$$

where P_c is the number of unique locator-bearing complete paths in the chapter, L_c is expanded locator occurrences, and W_c is indexable source words. The current calibration centers on 8 paths and 20 locators per 1,000 words, with target bands of 6-10 and 15-25 and broad tolerance bands of 4-12 and 10-30. Chapter ratings are aggregated by source-word weight. These are

framework-specific calibration values, not universal professional quotas; external validation is outstanding, and density contributes at most five points.

5.2 Deterministic six-dimension score

The rubric separates three layers that answer different questions: item-level diagnostic grades describe individual records; six dimension calculations summarize frozen ledgers; readiness gates restrict publication claims. Item grades are non-additive, and gates never add or subtract points. The dimension ratings are not selected by an evaluator. A declared calculation profile derives them deterministically from ledger statuses, raw numerators and denominators, structured defects, and uncertainty states.

Dimension	Points	Ledger-derived base calculation	Principal consequence caps
Meaningful coverage	20	Five times priority-weighted access; complete, partial, and missing receive 1, 0.5, and 0	Essential miss rate; critical central omission
Editorial selectivity	15	Substantive locator credit contributes 10 points; chapter-weighted density fit contributes 5	Systemic zero-credit locator patterns
Conceptual and stance fidelity	15	Mean node credit: pass 1, minor 0.85, major 0.55, fail 0	Major or critical meaning, relation, compound, or stance defects
Page-reference reliability	25	Five times F_R, combining strict locator support and treatment recall	High-value treatment misses; fabricated access; distributed unsupported patterns
Findability and navigation	20	Five times 60% coverage-conditioned task success, 30% architecture, and 10% reference validity	Failed tasks; destructive routes; recurrent architecture or reference defects
Mechanics and consistency	5	Mean node credit: pass 1, cosmetic 0.95, minor 0.85, major 0.55, fail 0	Incompleteness; critical, recurrent, or systematic defects
Total	**100**	**Sum of displayed dimension points**	**Weighted quality summary, not “percent correct”**

Substantive selectivity assigns credit 1 to substantive locator classes, 0.5 to mixed classes, and 0 to passing-mention, attribution, citation, example, and incidental classes; absent, unavailable, ambiguous, and out-of-scope records are routed separately. Reader tasks and references use 1, 0.5, and 0 for success, partial success, and failure; architecture uses the conceptual-credit map. A findability task is eligible only when every required subject has at least partial benchmark access, preventing a navigation score from double-counting a coverage failure as a successful route.

Consequence caps are applied to full-precision base ratings, never increase a rating, and may be triggered only by structured defect records that identify owner, severity basis, consequence, affected records or sections, recurrence, and applicable counts or rates. Free text cannot trigger a cap. Except for selectivity, the post-cap rating is rounded to the nearest half point using decimal ROUND_HALF_UP, and points are $r_d W_d/5$, rounded to two decimals. Selectivity rounds its substantive and density components separately to half steps before assigning their 10- and 5-point contributions; its displayed equivalent rating may therefore fall off a half step. Qualitative anchors are post-calculation reasonableness checks: a mismatch is reported but cannot manually change a rating.

Missing evidence is part of the calculation contract. In full mode, any required `not_measured` input blocks the affected dimension and total. A provisional score requires at least 95% of applicable units measured, subject to a declared small-denominator exception. `Uninspectable` units produce lower and upper endpoints; a numeric public rating is permitted only when both endpoints resolve to the same rounded rating and the same applied-cap identity. Otherwise the dimension is reported as insufficient evidence rather than forced to zero or silently dropped.

5.3 Critical gates

Publication readiness is a separate decision. The standard policy includes gates for out-of-scope or fabricated locators, systemic unsupported patterns, central omissions or misrepresentation, invalid compound headings, broken or substitutive cross-references, any third-level heading, systematic clutter, unresolved grounding, excessive uninspectable material, wrong source span, and incomplete output. A gate failure limits the claim that a candidate is publication ready but does not alter its score. This noncompensatory layer is necessary because a polished, comprehensive index can still fail publication-readiness criteria if it systematically directs readers to unsupported pages or omits a central conclusion.

6 Reproducibility, parallel review, and computation

6.1 Reproducible artifacts, not merely repeatable prompts

Every consequential artifact is schema validated and registered by relative path, SHA-256, visibility, retention class, and frozen status. The artifact manifest is written before shared state; changed identities invalidate downstream stages. Frozen objects are versioned rather than overwritten. A candidate-benchmark lock binds the candidate to the exact source, benchmark, policy, page map, chunks, audit mode, and uncertainty settings. The calculation artifact separately binds the calculation-profile identity and records every component’s raw numerator, denominator, normalized value, cap evaluation, uncertainty endpoints, post-cap rating, and point contribution. Two results are directly comparable only when both the evidence comparison key and calculation profile match.

The result references the calculation rather than restating unsupported arithmetic. A post-projection validation receipt cryptographically binds the calculation, result, and public report; verifies that public dimensions, total, gate outcomes, provenance, and observed or counterfactual score views agree; and fails closed on mismatch. Decimal values remain decimal through calculation and gate comparison rather than being converted to binary floating point; the regression suite includes the observed non-dyadic unsupported-locator rate of 0.086737.

Deterministic programs own page mapping, range expansion, identifiers, schema checks, hashing, routing, denominator validation, density and score arithmetic, diagnostic-grade derivation, checkpointing, and report projection. Human or model-assisted editorial review owns significance, substantive support, conceptual fidelity, stance, and navigation. The deterministic layer

can reproduce the arithmetic and trace every decision; it cannot make stochastic semantic judgments deterministic after the fact. Reproducibility here means preserving inputs, outputs, evidence, policies, denominators, cap provenance, uncertainty, and adjudications-not promising that a newly sampled model will independently recreate every judgment.

Restricted source PDFs, long quotations, raw candidate text, coordinates, recovery bundles, and private control records are separated from aggregate public evidence. Public reports use concise paraphrases and stable evidence identifiers. Cryptographic commitments make restricted inputs identifiable without redistributing copyrighted material.

6.2 Collision-safe parallelism

Parallel review changes throughput, not policy. Source-discovery workers own one immutable chapter artifact; locator workers own nonoverlapping page assignments; missing-access workers own deterministically routed subjects, treatment units, and tasks. Workers do not update shared state. A coordinator preflights an entire named batch before any merge, verifies its pull requests, performs one canonical integration, materializes private evidence, writes the integration receipt, updates the manifest, and updates state last; recovery procedures handle a partial external merge failure. Missing-access work begins only after every locator audit is canonically integrated and the evaluation validates; the structure audit and final score remain single whole-index operations.

This fan-out/fan-in design prevents lost updates, duplicate judgments, foreign page ownership, and denominator drift. It also makes failure recoverable at the artifact boundary rather than by reconstructing a long conversation.

6.3 Observed resource profile of the first application

Repository history records 51 merged pull requests: 16 in the benchmark repository-15 chapter-discovery proposals and one benchmark-QA/freeze proposal-and 35 in the candidate repository-one candidate-preparation, 17 locator-audit, and 17 missing-access proposals. `CHUNK-001` was later superseded by a candidate-blind v2 discovery artifact. Candidate audits were integrated in waves no wider than three. This is an observed orchestration width, not a measured number of simultaneous inference jobs.

Derived from repository commit timestamps, the observable interval from initialization of the candidate-blind benchmark repository to completion of the full source and candidate audit was approximately **4 days 21 hours**. The candidate-representation fidelity audit extended that observable interval to approximately **5 days 1 hour**. These wall-clock windows include agent orchestration, human decisions, pull-request review, queueing, checkpoint transfer, publication, and idle time. They are not processor-hours or model-runtime measurements.

At the pinned methodology commit, the evaluator exposes 76 JSON Schema contracts and passes 203 automated tests. A recorded validation run, workflow run 33022298105, completed its `validate` job in 33 seconds on GitHub's `ubuntu-latest` runner; the unit-test step occupied 21 seconds. This run predates some current contracts and is reported only as an observed lower-level validation workload, not a timing benchmark for semantic review. The workflow invokes no project-controlled GPU code, and hosted model hardware was not disclosed. Exact model snapshot, reasoning configuration, token use, inference calls, GPU type or hours, peak memory, energy consumption, human labor minutes, and billed cost were not recorded and are therefore not estimated.

The first application thus supports an artifact-scale and elapsed-workflow account, but not a hardware-normalized compute or cost claim. Future runs should emit an `execution-provenance` record for every stage containing model and version, reasoning setting, UTC start and comple-

tion times, input, cached and output tokens, tool calls, retries, branch and pull-request identifiers, wave membership, billed cost, human-review time, deterministic CPU time, and peak memory.

7 Worked application: The Oxford History of the French Revolution

The first full application illustrates the method; it is not yet a validation corpus. The frozen source benchmark covers the supplied 425-page body text of the 2002 edition of William Doyle's The Oxford History of the French Revolution. All supplied pages mapped one-to-one to printed labels 1-425. Front matter, notes or endnotes, bibliography, and publisher-index pages were absent and therefore excluded rather than treated as omissions. The candidate was a delivered 25-page, two-column index PDF from the IndexerLabs dataset, represented as the book's published index. The preparation report records complete internal reproduction; the candidate reference records authoritative-copy fidelity as `not_independently_verified`.

Evaluation object	Complete denominator
Source body pages / indexable words / chapter units	425 / 194,718 / 17
Frozen source subjects / relationships / reader tasks	1,366 / 3,460 / 1,026
Candidate heading nodes / displayed locator records	1,904 / 4,462
Atomic locator assignments / cross-references	5,338 / 16
Expected subject-treatment pages	3,210

The frozen evaluation result and its dimension calculations account for all 5,338 locator assignments, 1,366 source subjects, 3,210 expected treatments, 1,026 reader tasks, 1,904 heading nodes, and 16 cross-references. There were no duplicate or foreign worker assignments and no unresolved locators. The public projection is bound back to these artifacts by a validation receipt.

Seven localized, high-confidence conflicts between benchmark treatment classifications and canonical locator judgments were retained without reinterpretation for centralized adjudication; none constituted critical or major unresolved grounding.

Selected result	Observed value
Weighted six-dimension score	72.5 / 100
Strict supported-locator rate	4,643 / 5,338 = 86.98%
Binary locator precision	4,643 / 5,106 = 90.93%
At-least-partial locator support	4,875 / 5,338 = 91.33%
Expected treatment pages found	2,981 / 3,210 = 92.87%
Weighted concept access with partial credit	69.83%
Essential subjects missing	12 / 155 = 7.74%
Reader-task success, strict / partial credit	48.73% / 70.32%
Supported cross-references	13 / 16 = 81.25%
Standard readiness gates failed	7 / 13

Dimension	Rating	Points	Recorded applied cap
Meaningful coverage	3.5 / 5	14.0 / 20	Essential miss rate
Editorial selectivity	3.6667 / 5 equivalent	11.0 / 15	Systemic zero-credit pattern
Conceptual and stance fidelity	4.0 / 5	12.0 / 15	Major stance or relationship defect
Page-reference reliability	3.5 / 5	17.5 / 25	Distributed unsupported pattern
Findability and navigation	3.5 / 5	14.0 / 20	Destructive access route
Mechanics and consistency	4.0 / 5	4.0 / 5	Recurrent digit-for-accent substitution

Under the declared calculation profile, the supplied candidate fell in the 70-79 band: a useful foundation requiring substantial revision. The 72.5 total is not "72.5% correct" and is not asserted here as a score for an independently authenticated Oxford University Press index. Failed gates identified a systemic unsupported-locator pattern, central omissions, major stance errors, an unsupported compound relationship, an invalid substitutive `see`, unresolved cross-reference targets, and systematic clutter. Passing gates confirmed correct source span, in-scope and resolvable locators, no third-level headings, no excessive uninspectable material, no critical or major unresolved grounding, and complete structural accounting.

Candidate-fidelity review examined all 1,904 records, 4,462 displayed locators, 5,338 assignments, and 16 references. It found 14 records-18 character occurrences-with confirmed digit-for-accent substitutions; correcting the affected levée en masse entry also resolved one representation-caused cross-reference defect. A separately attributed representation-adjusted counterfactual scores 73.5/100. Because that view changes candidate evidence rather than evaluation rules, it does not replace the canonical as-delivered result. Both views remain in the same interpretive band; all seven failed gates and the `not_publication_ready` status remain unchanged.

The case also demonstrates why a human-created index should remain a candidate rather than the benchmark. The method can credit valid access absent from that index and can identify unsupported or missing access within it. It does **not** show that IndexPDF or AI indexing generally surpasses professional human indexing, nor does it establish a quality-adjusted cost advantage. Those are hypotheses for controlled comparative study.

8 Validity, limitations, and next tests

The method has strong internal auditability but several unresolved validity questions.

Benchmark completeness. Complete page inspection and a separate omission pass reduce omissions; they cannot prove that every valid access requirement was discovered. Results should report benchmark concept coverage, not exhaustive population recall.

Evaluator reliability. The first run did not measure inter-rater agreement, test-retest stability, or cross-model invariance. Evidence and adjudication make disagreement inspectable, but multiple independent evaluators are still required to estimate reliability.

Reader outcomes. Reader tasks in the benchmark are expert proxies. The method measures whether the index affords those tasks, not actual search success, time, comprehension, or

preference. A prospective user study should compare indexes using blinded factual, analytical, overview, and known-item tasks.

Calibration. Scope, policy, density targets, thresholds, and gates were frozen before candidate judgment; credit maps, component weights, cap thresholds, uncertainty rules, and rounding are bound by a versioned calculation profile before scoring. These choices have not been calibrated across genres. Cross-domain studies should test predictive validity and estimate sensitivity to weights, caps, and thresholds rather than presenting them as natural constants.

Generalizability. The first application covers one English-language scholarly history and one candidate. It does not establish performance on technical, medical, legal, trade, multilingual, illustrated, or electronic books.

Source and candidate fidelity. The Oxford source excluded matter absent from the supplied PDF, and the candidate was not independently authenticated against a publisher scan. Hashes establish which objects were evaluated, not whether those objects were the most authoritative possible copies.

Institutional independence. Candidate blindness prevents one important form of circularity, but the method was designed and first applied by the same project. Independent replication and professional-indexer review remain necessary.

Computation and cost. The first run did not log token use, exact model version, hardware, human minutes, energy, or cost. No economic comparison with human indexing is justified from repository timestamps.

A decisive comparative study would use multiple books and genres, at least two independently commissioned professional indexes per book, one or more machine-generated candidates, blinded candidate identities, a common frozen source benchmark, multiple independent evaluators, adjudication with agreement statistics, live-reader retrieval tasks, and prospective time, token, cost, and energy logging. Confidence intervals should accompany all aggregate differences. This design would test whether a candidate exceeds a professional index on source-grounded quality and reader performance at lower cost-without defining “better” as greater similarity to that professional index.

9 Conclusion

Finished subject indexes cannot be evaluated adequately by term overlap, rule compliance, locator existence, or visual inspection alone. A defensible evaluation must verify what the candidate asserts, search independently for what it omits, and judge whether the resulting access system works as an index. The methodology presented here does so by freezing a source-derived benchmark in contexts that never see candidate material before candidate judgment, auditing complete path-page claims in both directions, examining the global navigation system, preserving uncertainty and evidence, deriving dimensions deterministically from frozen ledgers, and separating the weighted quality score from noncompensatory publication gates.

Its principal benefit is epistemic: neither a human index nor an AI index receives authority from its provenance. Both are tested against the same source-grounded requirements, and every aggregate conclusion remains traceable through raw denominators, credit mappings, caps, score views, projection receipts, and contestable evidence. In the Oxford sensitivity analysis, the failed-gate set and publication-readiness conclusion remained stable under the confirmed representation-adjusted counterfactual. Broader replication, independent raters, reader studies, and prospective compute and cost telemetry are the next requirements.

Data and code availability

The method, schemas, deterministic helpers, and tests are available in the `evaluate-subject-index` repository at commit `31066101892d`. The frozen Oxford benchmark is available in `subject-index-benchmark-oxford-history-french-revolution-2002`, with the benchmark anchored at commit `98dbffd0ca17`. The candidate evaluation, canonical result, calculation artifact, counterfactual score view, web report, and validation receipt are available in `subject-index-evaluation-oxford-history-french-revolution-2002-original-published-index` at commit `0ffeca1f1a76`.

Copyright-restricted source and candidate PDFs, complete extracted text, private recovery artifacts, and the full item-assessment ledger are not redistributed. Public artifacts provide schemas, hashes, aggregate results, and source-safe evidence paraphrases.

Competing interests

John Camden created the evaluation methodology and develops IndexPDF, a separate subject-index generation system, through Publication Intelligence, LLC. The methodology is candidate agnostic, but its design and first application are not institutionally independent. The Oxford candidate was represented as a previously published human-created index and was not generated by IndexPDF; fidelity to an independently authoritative publisher copy was not verified.

References

Abdullah, N., and Gibb, F. (2008). “Using a Task-Based Approach in Evaluating the Usability of BoBIs in an E-book Environment.” In Advances in Information Retrieval, LNCS 4956, 246-257. https://doi.org/10.1007/978-3-540-78646-7_24.

Aït El Mekki, T., and Nazarenko, A. (2006). “An Application-Oriented Terminology Evaluation: The Case of Back-of-the Book Indexes.” In Proceedings of the LREC 2006 Workshop on Terminology Design: Quality Criteria and Evaluation Methods (TermEval), 18-21. Accessible copy: https://arxiv.org/abs/cs/0609133.

American Society for Indexing. (2015). “Best Practices for Indexing.” https://asindexing.org/best-indexing-practices/.

American Society for Indexing. (2025). “2025 ASI Indexing Awards.” https://asindexing.org/about/awards/asi-indexing-award/.

American Society for Indexing. (n.d.). “Index Evaluation Checklist.” https://asindexing.org/about-indexing/index-evaluation-checklist/.

American Society for Indexing AI Committee. (2026). AI and Book Indexing: Trajectory Data. White paper, March. https://asindexing.org/ai-news/ai-book-indexing-trajectory-data/.

Anderson, J. D. (1997). Guidelines for Indexes and Related Information Retrieval Devices. NISO TR02-1997. Bethesda, MD: NISO Press. https://www.niso.org/publications/tr02-1997-guidelines-indexes.

Bartmess, E., and Combs, M. R. (2025). “LLM-Generated Book Indexes: Can They Replace Professionally Created Indexes?” The Indexer, 43(4), 327-348. https://doi.org/10.3828/index.2025.33.

Bartmess, E., and Combs, M. R. (2026). “Can the Current Generation of Large Language Models (LLMs) Produce an Adequate Book Index?” The Indexer, 44(1), 35-48. https://doi.org/10.3828/index.2026.4.

Bennion, B. C. (1980). “Performance testing of a book and its index as an information retrieval system.” Journal of the American Society for Information Science, 31(4), 264-270. https://doi.org/10.1002/asi.4630310406.

Buckley, C., and Voorhees, E. M. (2004). “Retrieval Evaluation with Incomplete Information.” In Proceedings of SIGIR 2004, 25-32. https://doi.org/10.1145/1008992.1009000.

Coe, M. (2014). “Where is the evidence? A review of the literature on usability of book indexes.” The Indexer, 32(4), 161-168. https://doi.org/10.3828/indexer.2014.52.

Coe, M. (2015). “What do readers expect from book indexes and how do they use them? An exploratory user study.” The Indexer, 33(3), 90-101. https://doi.org/10.3828/indexer.2015.25.

Csomai, A., and Mihalcea, R. F. (2006). “Creating a testbed for the evaluation of automatically generated back-of-the-book indexes.” In Computational Linguistics and Intelligent Text Processing, LNCS 3878, 429-440. https://doi.org/10.1007/11671299_45.

Csomai, A., and Mihalcea, R. (2008). “Linguistically motivated features for enhanced back-of-the-book indexing.” In ACL-08: HLT, 932-940. https://aclanthology.org/P08-1106/.

Diodato, V. (1994). “User preferences for features in back of book indexes.” Journal of the American Society for Information Science, 45(7), 529-536. https://doi.org/10.1002/(SICI)1097-4571(199408)45:7%3C529::AID-ASI7%3E3.0.CO;2-O.

Diodato, V., and Gandt, G. (1991). “Back of book indexes and the characteristics of author and nonauthor indexing: Report of an exploratory study.” Journal of the American Society for Information Science, 42(5), 341-350. https://doi.org/10.1002/(SICI)1097-4571(199106)42:5%3C341::AID-ASI4%3E3.0.CO;2-7.

Doyle, W. (2002). The Oxford History of the French Revolution. 2nd ed. Oxford: Oxford University Press.

Gallina, Y., Boudin, F., and Daille, B. (2020). “Large-Scale Evaluation of Keyphrase Extraction Models.” In Proceedings of JCDL 2020, 271-278. https://doi.org/10.1145/3383583.3398517.

Golub, K., Soergel, D., Buchanan, G., Tudhope, D., Lykke, M., and Hiom, D. (2016). “A framework for evaluating automatic indexing or classification in the context of retrieval.” Journal of the Association for Information Science and Technology, 67(1), 3-16. https://doi.org/10.1002/asi.23600.

Gratch, B., Settel, B., and Atherton, P. (1978). “Characteristics of book indexes for subject retrieval in the humanities and social sciences.” The Indexer, 11(1), 14-23. https://doi.org/10.3828/indexer.1978.11.1.9.

Hurwitz, F. I. (1969). “A study of indexer consistency.” American Documentation, 20(1), 92-94. https://doi.org/10.1002/asi.4630200112.

IndexerLabs Team. (2026a). “Verifying 20,000 Index Locators at Scale.” April 13. https://indexerlabs.com/blog/verifying-20000-index-locators-at-scale (accessed August 28, 2026).

IndexerLabs Team. (2026b). “What We Learned Indexing the Same Book 120 Times.” April 12. https://indexerlabs.com/blog/what-we-learned-indexing-the-same-book-120-times (accessed August 28, 2026).

IndexerLabs. (n.d.). “Oxford History of the French Revolution Demo.” https://indexerlabs.com/oxford-history-of-the-french-revolution-demo (accessed August 28, 2026).

International Organization for Standardization. (1996). ISO 999:1996: Information and documentation-Guidelines for the content, organization and presentation of indexes. 2nd ed. https://www.iso.org/standard/5446.html.

International Organization for Standardization. (2026). ISO/FDIS 999: Information and Documentation-Guidelines for the Content, Organization and Presentation of Indexes. Final draft. https://www.iso.org/standard/87175.html.

Izzard, T. (2024). “Generative Artificial Intelligence (AI) and Its Performance at Indexing Tasks.” The Indexer, 42(4), 383-400. https://doi.org/10.3828/index.2024.24.

Järvelin, K., and Kekäläinen, J. (2002). “Cumulated Gain-Based Evaluation of IR Techniques.” ACM Transactions on Information Systems, 20(4), 422-446. https://doi.org/10.1145/582415.582418.

Johncocks, B. (2008). “Indexing by numbers: Is there scope for metrics in index evaluation?” The Indexer, 26(4), 158-162. https://doi.org/10.3828/indexer.2008.49.

Jörgensen, C., and Liddy, E. D. (1996). “Information access or information anxiety? An exploratory evaluation of book index features.” The Indexer, 20(2), 64-68. https://doi.org/10.3828/indexer.1996.20.2.3.

Markey, K. (1984). “Interindexer consistency tests: A literature review and report of a test of consistency in indexing visual materials.” Library & Information Science Research, 6(2), 155-177. https://eric.ed.gov/?id=EJ303179.

Marshall, L. (2023a). “Commissioning an Indexer (Part 3): How to Evaluate an Index-Guidance for Authors and Editors.” Society of Indexers, March 24. https://www.indexers.org.uk/posts/commissioning-an-indexer-part-3/.

Marshall, L. (2023b). “Qualities of a Good Index.” Society of Indexers, October 20. https://www.indexers.org.uk/posts/qualities-of-a-good-index/.

Medelyan, O., and Witten, I. H. (2006). “Measuring Inter-Indexer Consistency Using a Thesaurus.” In Proceedings of JCDL 2006, 274-275. https://doi.org/10.1145/1141753.1141816.

National Information Standards Organization. (2021). ANSI/NISO Z39.4-2021: Criteria for Indexes. https://doi.org/10.3789/ansi.niso.z39.4-2021.

Quinn, S. (2015). “Evaluating Indexes: Observations on ANZSI Experience.” The Indexer, 33(3), 107-112. https://doi.org/10.3828/indexer.2015.28.

Reich, P., and Biever, E. J. (1991). “Indexing Consistency: The Input/Output Function of Thesauri.” College & Research Libraries, 52(4), 336-342. https://doi.org/10.5860/crl_52_04_336.

Rolling, L. (1981). “Indexing consistency, quality and efficiency.” Information Processing & Management, 17(2), 69-76. https://doi.org/10.1016/0306-4573(81)90028-5.

Vagle, B. (2026). “AI Can Create Book Indexes: Here’s Why.” Indexia, May 28. https://www.indexia.tech/blog/indexing-standards-benchmark (accessed August 28, 2026).

Wittmann, C. (1990). “Subheadings in award-winning book indexes: A quantitative evaluation.” The Indexer, 17(1), 3-6. https://doi.org/10.3828/indexer.1990.17.1.3.

Wu, D., Yin, D., and Chang, K.-W. (2024). “KPEval: Towards fine-grained semantic-based keyphrase evaluation.” In Findings of ACL 2024, 1959-1981. https://doi.org/10.18653/v1/2024.findings-acl.117.

Wu, Z., Li, Z., Mitra, P., and Giles, C. L. (2013). “Can Back-of-the-Book Indexes Be Automatically Created?” In Proceedings of CIKM 2013, 1745-1750. https://doi.org/10.1145/2505515.2505627.