Migrating Off a Legacy ELN Without Losing Twenty Years of Data

What to migrate, what to preserve by other means, and what evidence to demand at each gate.

WHITE PAPER | SEPTEMBER 2026

PREPARED FOR

EXECUTIVES AND PROGRAM LEADS SIGNING AN ELN REPLACEMENT PLAN, WITH THEIR QUALITY, LEGAL, AND R&D DATA OWNERS

uncountable

Contents

WHO THIS IS FOR

This paper is written for the executive or program lead who has to sign an ELN replacement plan. It assumes the decision to replace has already survived the business case and is now stuck on risk. The decision it supports is narrow: what to migrate, what to preserve by other means, and what evidence to demand at each gate.

FOREWORD

One sentence ends the meeting

Almost every stalled ELN replacement stalls in the same place, and it is not the place the business case predicted. The cost model survives. The functional comparison survives. Then someone senior in the room says a version of the sentence that ends the meeting: we cannot risk losing twenty years of data. Nobody argues, because nobody can. The sentence is unanswerable in the form it is spoken, and so the project is deferred, and the legacy system acquires another year of records that will make the same conversation harder next time.

What makes that sentence so effective is that it sounds like one risk when it is really three, each with a different cause, a different test, and a different owner. Some of the fear is mechanical: records that do not arrive, attachments that arrive truncated, a degree symbol that turns into a question mark, a timestamp that shifts by five hours. That kind of loss is unpleasant but tractable, because it can be counted. Some of it is evidentiary: the record arrives, but the signature that made it defensible does not arrive as a signature, only as a sentence claiming that someone once signed. And some of it is semantic, which is the part that keeps people awake. A result of twenty-three, recorded in 2006, in a system whose units were implied by a template that no longer exists, described in shorthand invented by a group that has since been reorganized twice. Nothing in that record is corrupt. It simply no longer means anything to anyone who did not work there at the time.

Once the fear is split into those three parts, it stops being a veto and becomes a work plan. That is the whole purpose of this paper.

There is a second reframing that does at least as much work. Regulation does not require that every historical record end up inside the new system. It requires that records stay retrievable and readable for as long as they must be kept (21 CFR 11.10(c), EU GMP Annex 11 clause 7.1). That is a preservation duty, not a migration duty, and the difference between the two is most of the budget. A program that tries to move everything because moving everything feels safer is not being conservative. It is taking on the highest-risk version of the work, applying it uniformly to records nobody has opened since 2009, and calling the result diligence.

The third argument is borrowed rather than invented. Deciding whether a digital object is still the same object after its format has changed is not a new problem, and it was not first encountered in a laboratory. Archivists and national libraries have worked on it for thirty years and built a vocabulary for it: what information a future reader needs in order to interpret a file, who that future reader is assumed to be, which operations are reversible and which are one-way doors, and which properties must survive for an object to count as preserved at all. Almost none of that vocabulary appears in migration project plans, which is a shame, because it is precisely the language needed for an honest conversation about what will and will not survive.

The rest of this paper works through those ideas in order: what the obligations actually say, what transfers cleanly and what does not, where migrations break in practice, and what a signable, gated plan looks like. It also states, at the end, what it cannot settle. The intent is that a reader who has to put their name on this decision finishes it knowing which questions to ask, which answers to refuse, and what evidence to demand before releasing the next tranche of money.

0.1
What this paper argues

01 The three losses have three different owners. Mechanical loss belongs to data engineering, and is caught by reconciliation, hashing, and field comparison. Evidentiary loss belongs to quality and legal, and is usually solved by retention rather than transfer, because the binding between a record and its signature is enforced by the system that created it and no import routine recreates it. Semantic loss belongs to the scientists, cannot be tested by any reconciliation report, and is the reason most of these programs disappoint even when they technically succeed.

02 European GMP states the acceptance criterion for migration more precisely than most project plans do. Validation should include checks that data are “not altered in value and/or meaning during this migration process” (Annex 11 clause 4.8). Value is testable. Meaning is not, at least not by counting, and the meaning half of that sentence is where projects fail. This is also why the discipline worth borrowing from is digital preservation rather than data integration: the OAIS reference model’s ideas of Representation Information, a Designated Community whose knowledge changes over time, and non-reversible Transformation give a defensible way to decide what “the same record” means after a format change (CCSDS 650.0-M-2, published as ISO 14721).

03 The available evidence on outcomes justifies caution about schedule and cost. One practitioner study found that only 36 percent of migration projects kept to their original planned budget and 54 percent were delayed (Experian and Data Migration Pro research). The recurring cause is data quality discovered too late, which argues for profiling the legacy estate before committing to a scope. Two further warnings follow. Automated extraction of meaning from notebook prose is useful for triage and unreliable as a system of record: published named entity recognition on curated scientific text reaches an overall F1 around 87 percent, and the same authors note the model is “prone to making false-positive predictions” in production (Weston and colleagues, Journal of Chemical Information and Modeling, 2019). And extraction rights, export format specifications, and vendor assistance have to be settled while the incumbent still expects a renewal, because after notice is given the exit terms are whatever the contract already said. What should be signed, then, is a phased plan: tiered by retention obligation and by evidence of actual use, sequenced so that leverage is spent before notice is given, and gated on reconciliation evidence an inspector would accept.

Mechanical, evidentiary, semantic

Mechanical loss is byte-level and row-level failure. It is detectable and provable: record counts, checksums, and field-level comparison catch it, and the fix is engineering.

Evidentiary loss is subtler. The record arrives, but the thing that made it admissible does not. An electronic signature under US rules must be linked to its record so that it "cannot be excised, copied, or otherwise transferred to falsify an electronic record by ordinary means" (21 CFR 11.70).

> THE EVIDENTIARY TEST
>
> **A signature reconstructed as a text field in a new database is not that link. It is a claim about a link.**

Semantic loss is the one nobody can reconcile. A result of "23" is meaningless without the metadata that says milligrams, a point FDA makes directly in its data integrity guidance, which states that data should be maintained throughout the retention period "with all associated metadata required to reconstruct the CGMP activity" (FDA Data Integrity and Compliance With Drug CGMP, December 2018). Twenty years of notebook prose is full of site shorthand, superseded method names, and "see attached" pointers whose attachments were filed by a person who retired in 2011.

Naming which asset falls into which class turns an unbounded fear into three scoped work packages with three different acceptance tests.

MECHANICAL	EVIDENTIARY	SEMANTIC
Data engineering	**Quality and legal**	**The scientists**
Caught by reconciliation, hashing, and field comparison	Usually solved by retention rather than transfer	Cannot be tested by any reconciliation report

Fig. 1 One sentence, three risks. Each has a different owner and a different acceptance test.

Availability and readability, not residence

The obligations are about availability and readability over time, not residence in a particular application.

Part 11 requires procedures for "protection of records to enable their accurate and ready retrieval throughout the records retention period" and the ability to generate accurate and complete copies in both human readable and electronic form for agency inspection (21 CFR 11.10). Audit trails must be secure, computer generated, and time stamped, must not obscure previously recorded information, and must be retained at least as long as the records they document. Under GLP, raw data explicitly includes computer printouts and recorded data from automated instruments, archives must permit "orderly storage and expedient retrieval," and retention runs to the shortest of two years after application approval, five years after submission, or two years after study termination (21 CFR Part 58).

FDA's scope guidance is worth reading before assuming the worst about pre-2000 systems. The agency states it does not intend to enforce Part 11 requirements for validation, audit trail, record retention, and record copying as described in that guidance, and applies broader discretion to systems operational before 20 August 1997, while stressing that "records must still be maintained or submitted in accordance with the underlying predicate rules" (FDA, Part 11 Scope and Application, August 2003).

In the EU, stored data must be checked for accessibility, readability, and accuracy, and access must be ensured throughout the retention period (Annex 11 clause 7.1). The revision of GMP Chapter 4 under consultation is more explicit: records "should be available for review at any time during the required retention period, accessible in a human readable format," a documented disposal process must ensure originals or true copies are destroyed only after the retention period, and archiving is defined as long term retention of completed documentation and relevant metadata "in its final form for the purposes of reconstruction of a process or activity" (EU GMP Chapter 4 revision, documentation).

UK guidance is the most directly useful text here because it addresses legacy systems by name. It distinguishes archive, meaning protected data for long term storage, from backup, meaning "a copy of current (editable) data, metadata and system configuration settings maintained for recovery including disaster recovery," and states plainly that backups for recovery purposes do not replace long term retention in final form. It also warns that "the challenges of migrating data are often underestimated, particularly regarding maintaining the full meaning of the migrated records," and that where a system can no longer be supported, consideration should be given to maintaining the software for accessibility, possibly in a virtual environment (MHRA GXP Data Integrity Guidance and Definitions, revision 1, March 2018). Inspectorate expectations elsewhere align: PIC/S guidance for computerised systems lists archiving procedures, retrieval throughout the retention period, generation of accurate and complete copies, and secure time stamped audit trails as the issues to address wherever electronic records retain GxP data (PIC/S PI 011-3, 25 September 2007).

One further retention driver is often forgotten in R&D. Since the America Invents Act, US applications with an effective filing date on or after 16 March 2013 fall under provisions where an inventor's prior disclosure must be described "with sufficient detail and particularity" in a declaration (37 CFR 1.130). Derivation proceedings still turn on corroborated evidence: a showing "is not sufficient unless it is supported by substantial evidence, including at least one affidavit addressing communication of the derived invention and lack of authorization," and the communication showing "must be corroborated" (37 CFR 42.405). Notebooks no longer win priority races, but they still corroborate. That is a retention argument distinct from the GxP one and it should be documented separately.

CHAPTER 3: BORROW THE DISCIPLINE THAT ALREADY SOLVED THIS

Thirty years of preservation vocabulary

Archivists have spent thirty years on what it means for a digital object to survive a format change, and almost none of that vocabulary appears in migration project plans.

The OAIS reference model separates a Data Object from the Representation Information that "maps a Data Object into more meaningful concepts," and states that data interpreted using its Representation Information yields information. It defines a Designated Community whose knowledge base changes over time, and requires that preserved information be Independently Understandable, meaning interpretable "without having to resort to special resources not widely available, including named individuals" (CCSDS 650.0-M-2). That last clause is the entire ELN problem in one sentence. A 2004 formulation record whose abbreviations only the retired principal scientist can decode is, by this definition, already lost, whether or not it migrates.

OAIS also grades the operations. Refreshment and Replication copy without changing content. Repackaging alters only packaging. Transformation changes the Content Information or the Preservation Description Information and is described as generally a non-reversible transformation, verified against Transformational Information Properties whose preservation is "necessary but not sufficient" to show that information content survived. Reading a project plan through that taxonomy exposes which steps are reversible and which are one way doors.

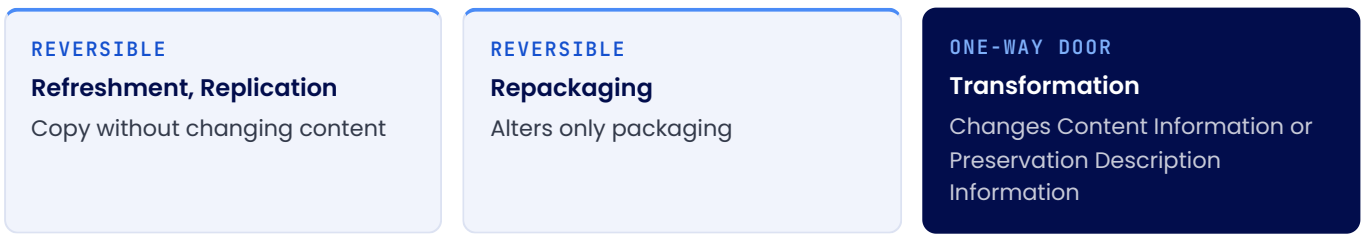

Fig. 2 The OAIS grading of operations. Reading a project plan through this taxonomy exposes which steps are reversible.

Two more concepts earn their place. PREMIS, the preservation metadata data dictionary maintained by the Library of Congress, models Objects, Events, Agents, and Rights, and is designed to support the "viability, renderability, understandability, authenticity, and identity of digital objects" (PREMIS Data Dictionary version 3.0). An Event with an Agent and an outcome is the structure needed to record that a given record was transformed from one format to another on a given date by a validated process, which is what an inspector will ask for. Significant properties, defined in preservation literature as the characteristics that must be preserved to ensure continued accessibility, usability, and meaning, are categorized as content, context, appearance, structure, and behaviour, with behaviour covering things like active links and updating calculations (InSPECT significant properties work, Knight on decoding significant properties).

Behaviour is the category that migration plans silently drop. FDA distinguishes a static record, such as an electronic image, from a dynamic one where "the record format allows interaction between the user and the record content" (FDA data integrity Q&A), and MHRA states that data must be retained in dynamic form where that is critical to integrity or later verification. Flattening a live calculation sheet to PDF is a decision to abandon a significant property, and it should appear in the plan as such rather than as an implementation detail.

Format choice follows the same logic, and it is worth making deliberately rather than inheriting it from whatever the export tool happens to produce. The Library of Congress sustainability factors, disclosure, adoption, transparency, self-documentation, external dependencies, impact of patents, and technical protection mechanisms, give a scoring frame for target formats, and the same guidance notes that deposit of full documentation in escrow with a trusted archive provides some degree of disclosure for proprietary formats (Library of Congress, Sustainability of Digital Formats). Obsolescence is a real risk, but proliferation of formats and versions is often the more immediate problem (Digital Preservation Coalition handbook on file formats). Where migration cannot preserve behaviour, emulation is the recognized alternative, though evaluating either approach against object properties "will only validate a necessary condition" for successful preservation (Guttenbrunner and Rauber on evaluating emulation and migration, CLIR overview of emulation as a preservation method). A conservative default from that literature is to retain the original bytestream in addition to any migrated derivative (CAMiLEON discussion of migration on request).

The artifact to negotiate first

The following table is the artifact to negotiate first. Nothing else in the plan can be scoped until each row has an agreed classification and a named owner.

ASSET CLASS	FIDELITY	WHAT PRESERVES IT IF IT DOES NOT TRANSFER	OWNER
Structured formulation and recipe records	Transfers cleanly if the schema maps; profile for dependencies and keys first (data profiling survey)	Field-level reconciliation report	Data engineering with R&D data owner
Test results with method and units	Transfers with transformation; unit and method mapping is the risk (FDA on metadata such as the unit "mg")	Documented unit conversion table, retained source values	Analytical or QC method owner
Instrument raw files (chromatography, spectra)	Often does not transfer with processing context; open interchange formats exist but are partial (JCAMP-DX, whose active development halted in 2006, AnIML)	Retain native files plus checksums; consider an archival container with an audit store (Allotrope Data Format)	Lab systems owner
Attachments, images, embedded OLE objects	Transfers with loss; embedded objects lose their editing application	Extract as discrete files, register format and dependency (LoC sustainability factors)	Data engineering
Free-text observations and local shorthand	Transfers as text, not as meaning	Glossary capture from incumbent staff; human verified extraction only	R&D knowledge owner
Audit trails	Rarely transfers with legal meaning intact	Archive in intelligible convertible form as required (Annex 11 clause 9)	Quality
Electronic signatures	Does not transfer as a signature; the record-signature link is system specific (21 CFR 11.70)	Read-only or archived source system as the evidentiary copy	Quality and legal
Workflow state and permissions	Does not transfer; semantics are application specific	Documented as-was configuration snapshot	System owner
Sample and inventory identifiers, cross-record links	Transfers with breakage; referential integrity fails where targets were purged	Identifier crosswalk table plus documented unresolved list	Data engineering

Fig. 3 Transfer fidelity by asset class. Every row needs an agreed classification and a named owner before scope can be set.

Mechanics, signatures, and meaning

5.1 Mechanics

The mechanical failures are the ones a competent team will fix, provided it goes looking for them early enough. Almost all of them are discovered by profiling rather than by testing, which is why the order of work matters more than the tooling.

Profiling comes before design. Profiling determines metadata about a dataset, from per-column null and distinct counts, data types, and value patterns up to uniqueness, functional dependencies, and inclusion dependencies across tables (Abedjan, Golab, and Naumann, The VLDB Journal). Those inclusion dependencies are how you find the links that will break. Identity matching across twenty years of user-entered sample names is a research problem in its own right, addressed in the literature by blocking and filtering techniques because naive comparison does not scale (survey of blocking and filtering for entity resolution).

Encoding is a high-frequency, low-glamour failure. Legacy code pages and Unicode transformation formats differ in byte order handling, and a byte order mark at the start of a stream signals encoding form rather than content (Unicode UTF-8 FAQ). Degree symbols, Greek letters in analyte names, and accented author names are where this shows up. Time is similar: civil time rules change by political decision and the reference database is updated periodically to reflect changes to boundaries, UTC offsets, and daylight saving rules (IANA Time Zone Database). A timestamp migrated without its zone rule may move.

Units deserve their own control. The canonical warning is not from pharma: the Mars Climate Orbiter mishap board traced loss of the spacecraft to ground software delivering impulse in pound-force seconds where newton seconds were specified, a factor of 4.45, with a resulting navigation disparity of roughly 169 kilometers (NASA Mars Climate Orbiter Mishap Investigation Board Phase I report). Every legacy result field without an explicit unit column is that failure waiting for a decimal point.

5.2

Signatures and audit trails

This is where plans quietly fail. A signature manifestation must show the printed name, the date and time, and the meaning of the signing, and must be included in any human readable form of the record (21 CFR 11.50). Signature to record binding is enforced by the originating system, and no import routine recreates the control environment that gave the original binding its meaning. The defensible position is to stop pretending. Migrate signature manifestations as attributed, non-repudiable historical metadata, state in the migration report that they are representations rather than re-executed signatures, and keep the source system or a validated archival package as the evidentiary copy for the retention period. Audit trails follow the same logic: EU guidance requires that they be available and convertible to a generally intelligible form, which is a preservation requirement that an export file can satisfy even when the new application cannot render them natively.

5.3

Semantic loss and what extraction can and cannot do

Free text is where most of twenty years of knowledge lives, and it is the least tractable asset. Automated extraction helps with triage. Chemical entity recognition benchmarks reached best F-scores of 87.39 percent for entity mentions and 88.20 percent for document indexing in the CHEMDNER evaluation (BioCreative IV CHEMDNER results), materials science NER trained on 800 hand-annotated abstracts reached an overall test F1 of 87.04 percent with entity normalization F1 of 94.5 percent (Weston and colleagues, 2019), and domain-specific polymer name extraction has been reported around F1 0.909 (SciNER). Those numbers are on curated published abstracts, not on twenty years of shift notes. Recall and precision trade off sharply in production settings, with reported configurations reaching recall of 0.95 at precision of 0.74 (chemical NER evaluation).

Large language models change the ergonomics, not the evidentiary standard. Evaluations of key information extraction in clinical text count both unsupported and contradicted facts across models (evaluation of LLM hallucination in medical extraction), and cross-model factuality benchmarks show wide and inconsistent performance (Hallucinations Leaderboard). An extraction that is right 90 percent of the time creates a specification file that is wrong 10 percent of the time and looks authoritative. Use extraction to build the search index, the glossary, and the candidate list for human review. Do not let extracted values become the approved value of a controlled record without a signed human check, and record that check as a preservation event.

Negotiate extraction before giving notice

Negotiate extraction before giving notice. Afterwards, the exit terms are whatever the contract already contained. Settle in advance: a documented export format specification rather than a promise of "CSV export"; extraction of attachments and binary payloads with their metadata; audit trail and signature history export with field definitions; a defined number of test extractions during the project; and an escrow or continuity arrangement covering the software and format documentation. The Library of Congress guidance on proprietary formats makes the archival case for escrowed documentation directly. In the EU there is now a statutory floor for cloud services: the Data Act requires that customers be able to switch between data processing services, with the scope of exportable data including input data, output data, and metadata directly or indirectly generated by the customer's use of the service, and applies generally from 12 September 2025 (Regulation (EU) 2023/2854).

A "CSV export" typically omits everything that matters here: attachment binaries, audit trail rows, signature manifestations, workflow state, template versions, permission history, and the referential links between records. Ask for a sample extract of one hundred real records across the worst five templates you own, and reconcile it before signing anything.

Tiering makes the scope defensible. Access frequency evidence from the incumbent system, meaning which records were opened, searched, or cited in the last five years, is the strongest available basis for tier assignment, and it converts an argument about sentiment into an argument about evidence.

TIER	CONTENTS	DRIVING OBLIGATION	TREATMENT
Active migration	Records in current use: live specs, current methods, open studies	Business continuity plus predicate rule availability (21 CFR 11.10)	Full structured migration with reconciliation
Reference migration	Historically consulted records, superseded formulations, closed studies still searched	Readability throughout retention (Annex 11 clause 7.1)	Migrate content and key metadata; accept documented loss of behaviour
Archival only	Raw instrument files, audit trails, signature history, GLP raw data	Archive and retrieval duties (21 CFR 58.190, MHRA archive versus backup)	Preservation package plus read-only or virtualized source system
Legally destroyable	Records past all retention periods with no patent or litigation hold	Documented disposal process (EU GMP Chapter 4 revision)	Controlled destruction with records of what was destroyed

Fig. 4 Four tiers, assigned by retention obligation and by evidence of actual use.

Keeping the old system alive as a read-only archive is a legitimate tier three treatment and a real liability. Its operating system will reach end of support on a published date, as with the extended support end date of 10 October 2023 for one widely deployed server platform (Microsoft lifecycle), and continued use of end-of-life software "poses consequential risk" that an attacker can exploit (CISA on end-of-life software). Read-only retention must therefore come with network isolation, a licensing position, and a dated exit plan of its own.

CHAPTER 7: STRATEGY, EVIDENCE, AND PROOF

Tranches, not a single weekend

Big-bang cutover concentrates all risk on one weekend and assumes the target behavior is fully specified. The incremental alternative, described in software modernization as the strangler fig approach, moves behavior gradually from the legacy code base to the new one, on the grounds that wholesale replacement "has gone down in flames most of the time" and that it is hard to figure out the details of existing behavior (Martin Fowler on the strangler fig application). The practical ELN analogue is: new work starts in the new system on day one, tier one content migrates in tranches by project or site, tier two migrates on demand as records are requested, and tier three never migrates at all.

It is worth being candid with a steering committee about what the published outcome data does and does not support. The survey evidence supports caution rather than confidence. One practitioner study found only 36 percent of migration projects kept to the original planned budget and 54 percent suffered delay (Experian and Data Migration Pro); an analyst figure widely repeated in that literature puts the share that fail or exceed budgets and timelines at 83 percent (Experian citing Gartner). Separate analyst survey work reported average savings of about 170,000 dollars per project from applying its identified success factors (Bloor Research data migration survey 2011). These are trade sources, useful for justifying contingency, not for forecasting.

Validation of the migration is a separate deliverable from validation of the new system. Industry guidance treats data migration as its own topic with a dedicated appendix in the current computerized systems framework (GAMP 5 second edition, appendix D7 on data migration). The evidence set that survives inspection is unglamorous: record counts by object type reconciled source to target; cryptographic hashes over extracted binaries, since any change to a message will with very high probability produce a different digest (FIPS 180-4, Secure Hash Standard); field-level comparison on a statistically justified sample using a published attribute sampling scheme indexed by acceptance quality limit (ISO 2859-1); documented unit and code list mappings; and a preservation event log recording who transformed what, when, and with what outcome, in the PREMIS sense.

PHASE	OBJECTIVE AND ENTRY CRITERIA	EXIT CRITERIA	DURATION	FAILURE MODE PREVENTED
0. Obligation mapping	Establish what must remain readable and for how long; entry is a named quality and legal owner	Signed retention register mapping each record class to obligation and end date (21 CFR 58.195)	4 to 8 weeks	Migrating everything because nobody knows what may be retired
1. Profiling and fidelity classification	Profile the legacy estate and classify every asset class; entry is read access plus access-frequency logs	Completed transfer-fidelity table and tier assignment; documented broken-link inventory (profiling survey)	6 to 12 weeks	Discovering data quality after contracts are signed
2. Extraction rights and test extract	Secure format specs and extraction support; entry is that notice has not yet been given	Successful sample extract across worst-case templates, including attachments and audit trails	4 to 10 weeks, overlapping	Losing leverage and finding out CSV omits the evidence
3. Preservation package build	Build archival packages for tier three; entry is agreed target formats	Fixity-verified packages with representation information and preservation events (OAIS, PREMIS)	8 to 16 weeks	Archive that is really just a backup
4. Pilot migration	Migrate one site or product family end to end; entry is approved migration plan and rollback	Reconciliation pack passes: counts, hashes, sampled field comparison	8 to 12 weeks	Scaling a broken transform across the whole estate
5. Parallel operation	Run old read-only and new authoritative; entry is pilot acceptance	Defined period with no unresolved discrepancies and search parity accepted by users	One to two full business cycles	Cutover before anyone can find anything
6. Tranche migration and closeout	Migrate remaining tiers; entry is stable pilot	Final reconciliation, decommissioning record, archive access procedure tested	6 to 18 months	Indefinite dual running and unbudgeted legacy costs

Fig. 5 The phased sequence, each phase named by the failure mode it prevents.

Four gates, each with named evidence

Approve the sequence above with four explicit go/no-go gates, each with named evidence.

Gate A, after phase 1.	Evidence: the retention register, the transfer-fidelity table with every row owned, tier volumes, and the broken-link inventory. Decision: approve scope and budget for extraction, or re-scope.
Gate B, after phase 2.	Evidence: the test extract reconciliation showing what the incumbent export does and does not contain, plus signed extraction terms. Decision: proceed to build, or renegotiate before giving notice.
Gate C, after phase 4.	Evidence: pilot reconciliation pack with record counts by object type, hash verification for binaries, sampled field-level comparison against the agreed acceptance quality limit, the unit and code mapping document, and a written statement of what was transformed non-reversibly and what evidentiary artifacts remain in the source system. Decision: approve tranche rollout.
Gate D, before decommissioning.	Evidence: successful retrieval test performed by quality against the archival packages, without help from the migration team, plus the documented disposal record for anything destroyed. Decision: release the legacy environment, or extend read-only retention with a funded plan.

Fig. 6 The four go/no-go gates, with the evidence each one requires and the decision it carries.

Three conditions should stop the program and trigger re-scoping rather than more effort. First, if profiling shows that the identifier crosswalk cannot resolve a material share of cross-record links, because migration would then propagate silent breakage into the system of record. Second, if the incumbent cannot or will not supply audit trail and signature history in a defined format, because the evidentiary tier then has to be solved by archival retention before any cutover date can be credible. Third, if the pilot reconciliation fails on meaning rather than on counts, meaning values arrive intact but their units, methods, or statuses are misinterpreted, because that is the failure Annex 11 clause 4.8 exists to catch and it cannot be fixed by rerunning the load.

Set leadership expectations accordingly. The honest shape of this program is measured in quarters for tiers one and two and in years for full decommissioning, and historical search will get worse before it gets better, because twenty years of accumulated shorthand does not index well on day one. Saying that at the start is cheaper than discovering it at Gate C.

None of this makes the original fear unreasonable.

THE REAL FAILURE MODE

Twenty years of records really can be lost, and the most common way to lose them is not a failed load script but a successful one, run over data whose meaning nobody captured while the people who held it were still available to ask.

The reframing this paper offers is simply that the risk is specific, and specific risks can be scoped, owned, evidenced, and signed. A plan that names its three losses, tiers its estate by obligation and use, spends its leverage before giving notice, and refuses to advance without reconciliation evidence is not a guarantee. It is, however, a defensible answer to the sentence that has been ending the meeting.

CHAPTER 9: WHAT THIS PAPER DOES NOT SETTLE

Five open questions

1. The retention register is jurisdiction and product specific. The obligations cited here are the common GxP and patent drivers, not a complete legal analysis, and litigation holds or local law may extend any of them.

2. The evidentiary standing of migrated signature manifestations has not been tested uniformly across inspectorates. The conservative treatment recommended here, retaining the source system or an archival package as the evidentiary copy, is a risk position rather than a settled requirement.

3. Published extraction accuracy figures come from curated scientific corpora. There is no comparable published benchmark for twenty years of internal notebook prose, so the human verification burden cannot be estimated from these numbers alone.

4. The migration failure statistics available in retrievable sources are trade research with undisclosed sampling. They justify contingency but should not be quoted as base rates.

5. Migration and emulation are not fully substitutable, and the preservation literature is explicit that verifying preserved object properties validates only a necessary condition. Where behaviour is the significant property, no approach here fully removes the risk of losing it.

uncountable

WHITE PAPER

The risk is specific. Specific risks can be scoped, owned, evidenced, and signed.

Name the three losses, tier the estate by obligation and use, spend the leverage before giving notice, and refuse to advance without reconciliation evidence.

ABOUT UNCOUNTABLE

Uncountable connects R&D, quality, product lifecycle, and project management on one structured data model. Our founding team of MIT and Stanford engineers brings expertise in software development and data analytics.

REQUEST A DEMO

SCAN TO BOOK

REFERENCES

01 **21 CFR Part 11,** Electronic Records; Electronic Signatures.

02 **21 CFR Part 58,** Good Laboratory Practice for Nonclinical Laboratory Studies.

03 **FDA.** Part 11, Electronic Records; Electronic Signatures — Scope and Application. August 2003.

04 **FDA.** Data Integrity and Compliance With Drug CGMP: Questions and Answers. December 2018.

05 **EudraLex Volume 4, Annex 11,** Computerised Systems.

06 **EU GMP Chapter 4,** Documentation (revision under consultation).

07 **MHRA.** GXP Data Integrity Guidance and Definitions, revision 1. March 2018.

08 **PIC/S PI 011-3,** Good Practices for Computerised Systems in Regulated GxP Environments. 25 September 2007.

09 **CCSDS 650.0-M-2,** Reference Model for an Open Archival Information System (OAIS), published as ISO 14721.

10 **PREMIS Data Dictionary** for Preservation Metadata, version 3.0. Library of Congress.

11 **Library of Congress.** Sustainability of Digital Formats.

12 **ISPE.** GAMP 5, second edition, appendix D7 on data migration.

13 **FIPS 180-4,** Secure Hash Standard, and **ISO 2859-1,** sampling procedures for inspection by attributes.

14 **Regulation (EU) 2023/2854** (Data Act), on harmonised rules on fair access to and use of data.