EARTHSCAN

WHITEPAPER 20 PAGES / MAY 2025

Recovering Trapped Data: A Five-Year View of Industrial Digitisation

Look back across five years of raster-to-vector work and the individual projects stop looking like separate projects. The same shape recurs every time: a huge archive of scanned documents nobody can query, a synthetic-data bootstrap that manufactures the training set the archive never had, a learned segmenter that turns paper into trustworthy tabular data, and a marginal cost per recovered document that keeps falling because one trained model amortises across the whole corpus. This whitepaper reads that path at portfolio altitude, using the sourced numbers from a reference engagement, and argues that the repeatability is the asset, not any one model.

The EarthScan Team
Research, engineering, and field

INDUSTRIAL-DIGITISATION / RASTER-TO-VECTOR / TRAPPED-DATA / SYNTHETIC-DATA / DOMAIN-RANDOMISATION / SEMANTIC-SEGMENTATION

CONTENTS

> "
>
> Five separate projects, seen from far enough away, are one project done five times. The trapped archive is always the starting condition, the synthetic bootstrap is always how training becomes possible, the learned segmenter is always what turns paper into data, and the cost of the next document always falls. The repeatable path is the asset.
>
> "

I

THE PATTERN

Why the individual projects stop looking separate

The shape that recurs

When you stand close to a single raster-log digitisation engagement, it looks bespoke. This operator, this archive, these curve types, this delivery deadline. Stand back across five years of the work and the bespoke detail falls away, and what is left is a shape that recurs with almost no variation. There is always a large archive of scanned documents that nobody can query. There is never a labelled training set for it, and there is never a realistic way to produce one by hand at the scale the archive demands. So the training set is manufactured synthetically instead. A learned segmenter is trained on that synthetic set and then pointed at the real archive, where it turns pixels back into numbers. And once the model exists, the cost of recovering one more document falls, because a single trained model serves the whole corpus rather than being rebuilt for each item in it.

This whitepaper is the portfolio-altitude reading of that shape. It is deliberately not a re-telling of any single project. We have written elsewhere about the architecture that does the segmentation, VeerNet, and about the sub-pixel validation that decides whether a recovered curve is trustworthy. Those are close-up documents. This one is the wide shot: the claim that the value in five years of raster-to-vector work is not any one model we trained but the repeatability of the path itself, and that the path has four stages that always arrive in the same order.

We ground the wide shot in the sourced numbers from one reference engagement, a digitisation programme for a Texas onshore operator, because concrete figures keep a retrospective honest. The archive we worked was 136,771 TIF files plus 7,781 LAS files. The training set we manufactured was 15,000 multiclass and 2,000 binary synthetic instances. The segmenter reached a peak coefficient of determination of 0.9891 and a peak curve-

class intersection-over-union of 0.51. The accelerated build ran 16 weeks, served on a GPU that rents for 750 to 1,800 EUR per month. Every one of those numbers belongs to one project. The argument is that the pattern they sit inside belongs to all of them.

What "trapped" actually means

The word trapped is doing real work, so it is worth being precise about it. A scanned well log is not missing data. It contains every number the original paper log contained: the gamma-ray value at every depth, the resistivity, the caliper, the depth scale itself. What it lacks is queryability. The numbers are present as the arrangement of dark and light pixels that a plotter once drew, and no calculation, no join against a well database, no search, no petrophysical model can reach them while they stay in that form. The information is there and it is inert. Recovering it means converting the picture of the numbers back into the numbers, which is the raster-to-vector problem stated in one sentence.

That is why digitising legacy archives is worth doing at all, and why the payoff is disproportionate to the effort it looks like it should require. The subsurface literature has been consistent that the industry's records are its underused asset and that turning them into machine-readable form is a precondition for almost everything else, from reservoir characterisation to the acceleration of interpretation workflows that run orders of magnitude faster than manual mapping once the inputs are digital [1]. An archive of 136,771 scanned images is not a storage problem. It is a hundred thousand documents of production-relevant data that no model, no query, and no analyst can use until the raster becomes a vector.

II

STAGE ONE

The starting condition is always a corpus nobody can query

The archive is the problem statement, not the input

In most machine-learning projects the dataset is the input: you begin with labelled data and design a model to fit it. Industrial digitisation inverts that. The archive is not the input to the work; it is the problem statement. It defines the scale, it defines the variety the model has to survive, and it defines the fact that no supervised training is possible against it directly, because not one of its documents carries a label that says which pixels are signal and which are background.

The scale is the first thing the archive dictates. 136,771 TIF files is not a number you annotate by hand. A senior interpreter working a raster log clicks along each curve at depth intervals and calibrates the depth scale, and that is a matter of one to a few hours per log on the manual workflow the industry has lived with for decades. Multiply even the optimistic end of that by a hundred thousand documents and the manual route is not slow, it is arithmetically impossible on any human timescale or budget. The size of the corpus is the reason the work has to be learned rather than performed.

The variety is the second thing it dictates. An archive that deep is not uniform. It holds different vintages of paper, different plotter conventions, different curve sets, photocopies of photocopies, folds, stains, and skew. The 7,781 LAS files that sit alongside the rasters are a reminder that the corpus is mixed: some of the well's data already exists in a digital, queryable form, which is exactly what makes it valuable as a check on the recovered curves, but the vast majority of the archive is the rasters, and the rasters are the part with no labels. The model that recovers this corpus has to survive its full range, not a clean subset of it.

STAGE ONE: THE TRAPPED ARCHIVE THAT SETS EVERY DOWNSTREAM DECISION

136,771

Scanned TIF files, the corpus to be recovered

7,781

LAS files already digital, useful as a cross-check

0

the constraint

Hand labels available for supervised training

1-4 hrs

Manual tracing per log, the route that does not scale

The lesson of stage one, repeated across every engagement, is that the archive is not something you receive and start using. It is something that hands you a constraint on day one: a corpus too large to label and too varied to shortcut, with no supervised signal in it. Everything the next stage does is a response to that constraint.

Why the archive is the problem statement and not the input, expressed as one multiplication. Take the sourced corpus of 136,771 scanned TIF files and multiply by the sourced manual-tracing budget of one to four hours per log, and the total is a wall of interpreter-hours; divide by a working year and it becomes interpreter-years, plotted on the left bar. Against it stands the learned build the engagement actually ran, a 16-week accelerated track with a 6-person team, plotted on the right bar in the same calendar-year unit. Drag the manual-rate lever anywhere across the sourced one-to-four-hour band and the manual wall never comes down to meet the learned build; the readout at top right is the ratio between them. The orange bar is the only element that argues: it is the manual route the instrument shows to be impossible, which is exactly why the corpus forces the work to be learned rather than performed. The corpus size, the one-to-four-hour manual band, the 16-week timeline, and the 6-person team are sourced from the engagement archive and the whitepaper's stage-one narrative; the working-hours-per-year figure is an illustrative divisor used only to convert interpreter-hours into calendar-years, and no cost or price is asserted.

The wall above is what "too large to label" means in arithmetic rather than adjective. Multiply the sourced corpus by any rate in the sourced one-to-four-hour band and the hand-tracing route is not a slower version of the automated one; it is a different order of magnitude of calendar-time, and the ratio to the learned build is the size of the problem the rest of the path exists to solve.

III

STAGE TWO

You cannot annotate your way out of an annotation problem

The synthetic bootstrap manufactures the dataset the archive never had

The constraint from stage one has an escape that is not obvious the first time you meet it and becomes the default once you have. If the real archive has no labels and cannot practically be given them, do not try to label the archive. Manufacture a synthetic archive instead, one where the label is free because you drew the image and therefore know exactly which pixels are which. Generate a well log image and, in the same pass, generate the pixel-perfect mask that says this run of pixels is curve one, that run is curve two, the rest is background. Now you have supervised training data for a task that had none, and you produced it without a single human annotation of the real corpus.

For the reference engagement the manufactured set was 15,000 multiclass instances and 2,000 binary instances. Those are not photographs of real logs; they are procedurally generated logs, drawn by a generator whose parameters are swept across the variety the real archive threatens to throw at the model. The reason this works, and the reason it is not a shortcut that quietly fails on real data, is domain randomisation: if the generator's nuisance parameters, the line weights, the noise, the grid styles, the skew, the contrast, are randomised widely enough, the real archive stops looking like an unseen domain and starts looking like one more sample from the synthetic distribution the model already trained on [2]. The goal of the generator is not to make one beautiful realistic log. It is to make fifteen thousand different-enough logs that the real ones fall inside their spread.

This is the stage that most decisively separates industrial digitisation from a standard supervised-learning project, and it is the stage that makes the whole path repeatable. A generator is a reusable asset in a way a hand-labelled dataset is not. Once you can

synthesise labelled logs for one operator, synthesising them for the next operator's curve set is a parameter change, not a fresh annotation campaign. The training data is built, not collected, and built things can be rebuilt cheaply.

> **WHY THE BOOTSTRAP IS THE PIVOT OF THE WHOLE PATH**
>
> The trapped archive gives you a task with no training data. The synthetic bootstrap gives you training data with no annotation. That single move, manufacturing the labels instead of harvesting them, is what turns an impossible supervised problem into a routine one, and it is why the same path applies to any archive of plotted documents, not just well logs. Randomise the generator wide, not realistic; the real corpus has to land inside the synthetic spread, not next to one perfect example of it.

The generator is where the domain knowledge lives

It is tempting to think the intelligence of the system lives in the model. Across five years the more accurate reading is that a great deal of it lives in the generator. The generator encodes what a well log is: how many curves, how they cross, what the grid looks like, how the depth scale runs, what degradation real scans suffer. Getting the generator right is getting the domain right, and a generator that spans the real variety is worth more than a marginally better network trained on a generator that does not. This is why the bootstrap is not a preliminary step to be rushed through on the way to the model. It is the step where the project's understanding of the domain is deposited, and the model is downstream of it.

IV

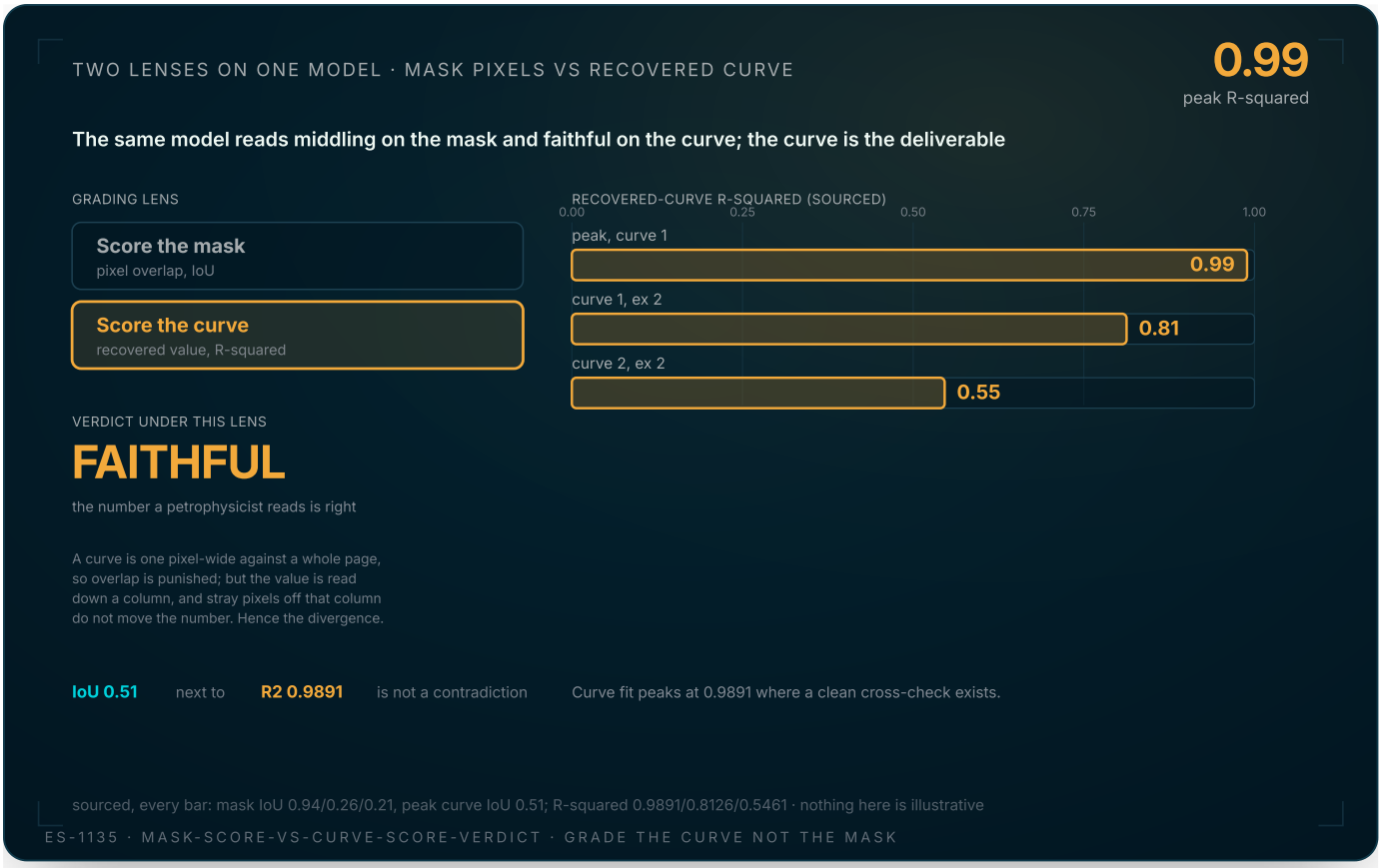
STAGE THREE

The learned segmenter turns the picture back into the numbers

From pixels to a mask to a curve

With a manufactured training set in hand, stage three is the one that looks most like conventional machine learning, and it is. A segmentation network in the encoder-decoder family consumes the scanned image and emits a per-pixel classification: this pixel is curve one, this is curve two, this is background [3]. The encoder compresses the image into features, the decoder expands those features back to full resolution, and skip connections carry the fine detail from the early layers across to the late ones so the thin curve traces survive the round trip. On the reference engagement the network also carried an attention refinement on the bottleneck, but the architectural detail is the subject of the close-up documents. At portfolio altitude the point is narrower: this stage exists to convert the picture into a mask, and then the mask into an ordered sequence of depth-indexed numbers, which is the vector the whole exercise was after.

What matters at this altitude is how you decide the segmenter is good enough, because the answer is not the one a segmentation benchmark would give. The pixel-overlap metric the field reaches for first, intersection-over-union, peaked at 0.51 on the curve classes for this engagement. Read as a segmentation score that looks middling. Read correctly it is not the score that matters, because the deliverable is not the mask, it is the recovered curve, and the right question is how well the reconstructed curve agrees with the true one. On that measure the segmenter reached a peak coefficient of determination of 0.9891, meaning the recovered curve tracks the reference almost exactly where a clean cross-check exists. The gap between an IoU of 0.51 and an R-squared of 0.9891 is not a contradiction. It is the whole reason we grade the digitiser on the curve and not on the mask: a model can be imperfect at the pixel level and still recover a curve that is faithful to the number a petrophysicist needs, because a curve is a one-dimensional object read down a column of pixels, and small pixel errors that do not move the column position do not move the value.



Why the digitiser is graded on the recovered curve and not on the pixel mask, shown as one model read through two lenses. Toggle the grading lens on the left. Under the mask lens the plotted bars are the sourced per-class intersection-over-union: background overlaps almost perfectly at 0.94, but the thin curve classes score 0.26 and 0.21 and the peak curve-class IoU is 0.51, because a curve is a few pixels wide against a whole page of background and every stray pixel is punished. Under the curve lens the bars are the sourced coefficient of determination between the recovered curve and the reference, which peaks at 0.9891 where a clean cross-check exists, because the value is read down a column and small pixel errors that do not move the column position do not move the number. The verdict headline flips from MIDDLING to FAITHFUL on the same model as the lens changes, which is the whole point: an IoU of 0.51 sitting next to an R-squared of 0.9891 is not a contradiction. The orange lens is the curve score, the only one that is the deliverable. Every bar on this instrument is a sourced figure from the engagement archive; nothing here is illustrative.

Toggle the lens above and the same model earns two different verdicts on the same sourced numbers: middling when the pixel mask is the yardstick, faithful when the recovered curve is. Only the second yardstick measures the thing we ship, which is why every accuracy claim in this whitepaper is a curve claim and not a mask claim.

STAGE THREE: WHAT THE LEARNED SEGMENTER ACTUALLY DELIVERS

0.9891
Peak R-squared, recovered curve vs reference

0.51
Peak curve-class IoU, the pixel score that is not the point

15,000
Multiclass synthetic instances the model trained on

2,000
Binary synthetic instances for the two-class case

This is also the stage where the honest limit of the work is set. The model is fast and accurate and not perfect. A peak curve-class IoU of 0.51 means it is excellent on most of the trace and wrong on a minority of it, and the recovered corpus is only trustworthy if there is a way to see and correct that minority. Across engagements the constant is that the segmenter is never shipped as an unattended oracle. It is shipped as the thing that does the ninety-plus percent of the tracing that used to consume an interpreter, with a review step that puts a human exactly on the part the model got wrong. The value is in the ratio: the model moves the human from doing all of the work to checking a small fraction of it.

V

STAGE FOUR

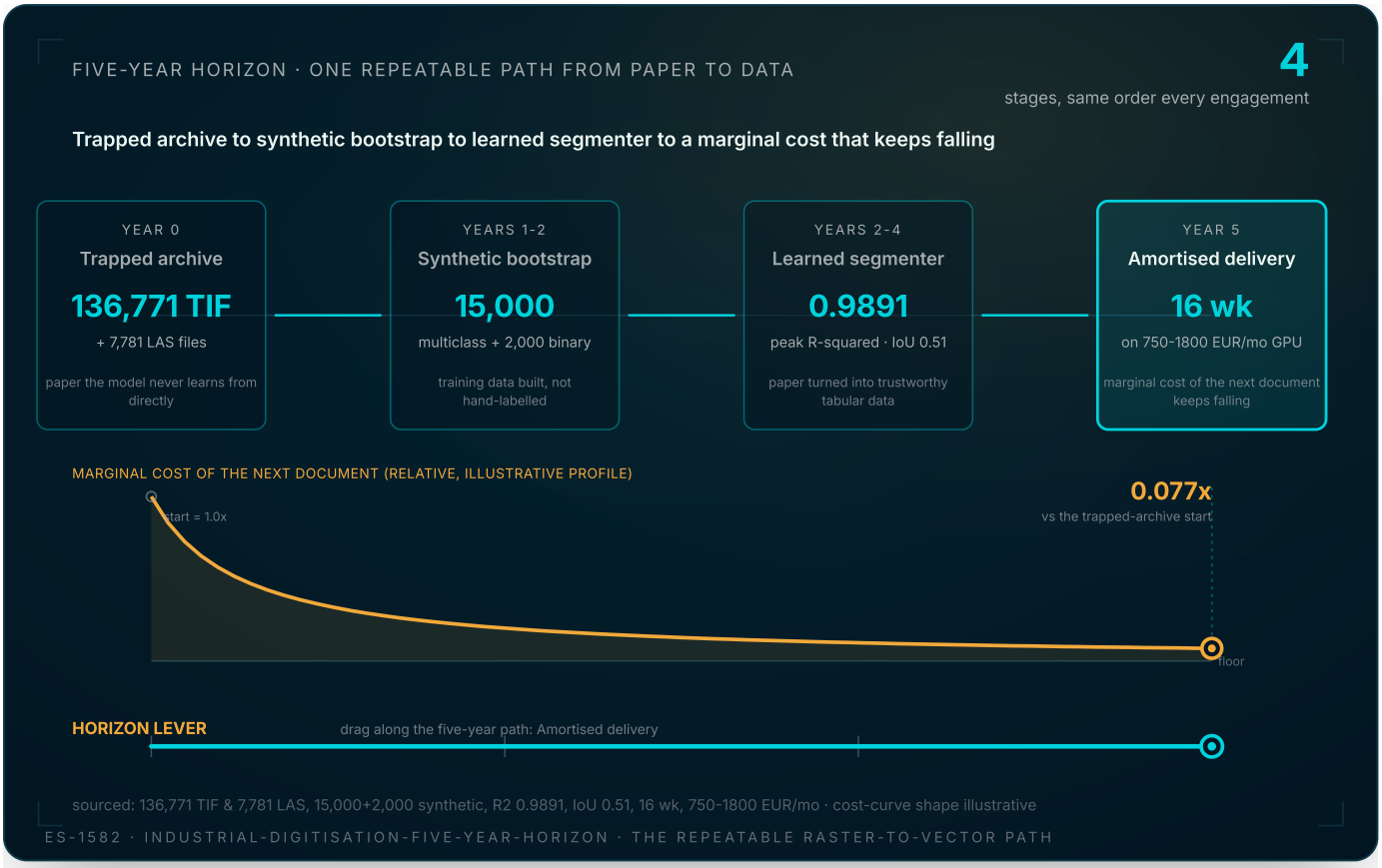
The cost of the next document keeps falling

Training is a one-off; recovery is not

The fourth stage is the one that makes the retrospective worth writing, because it is the stage where the economics of the path reveal themselves. Training the segmenter is a one-off. It happens once per build, it costs what it costs, and on the reference engagement the accelerated build ran 16 weeks on a GPU that rents for 750 to 1,800 EUR per month depending on tier. That is the fixed cost. Recovering documents with the trained model is the recurring activity, and its cost per document is dominated by serving compute, not by anything that scales with the corpus the way manual tracing did. So the marginal cost, the cost of recovering one more document once the model exists, falls as the corpus grows, because the fixed build is spread across ever more recovered documents.

This is not an accounting trick and it is not unique to well logs. It is the general property of any system where a model is trained once and served many times, and it is the reason the trained model should be understood as a small box inside a large system rather than the system itself [4]. The model is the part that took the research. The falling marginal cost is a property of the serving system that surrounds it: one trained model, pointed at a hundred thousand documents, amortising a 16-week build across all of them. The more documents the corpus holds, the cheaper each recovered document becomes, and an archive of 136,771 files is a very large denominator to divide a one-off build across.

The instrument below reads the whole path as one picture, and it is where the argument peaks.



The repeatable raster-to-vector path at portfolio altitude, read left to right across a five-year horizon. Stage one is the trapped archive, 136,771 TIF files and 7,781 LAS files that the model never learns from directly. Stage two is the synthetic bootstrap, 15,000 multiclass and 2,000 binary training instances built rather than hand-labelled. Stage three is the learned segmenter, which reaches a peak R-squared of 0.9891 and a peak curve-class IoU of 0.51 and turns paper into trustworthy tabular data. Stage four is amortised delivery, a 16-week accelerated build served on a 750 to 1,800 EUR per month GPU. Drag the horizon lever to stand at any stage; the one orange element, the falling marginal-cost curve underneath, is the argument: the cost of digitising the next document collapses toward a floor as one trained model amortises across the whole corpus. The archive counts, the synthetic instance counts, the R-squared, the IoU, the 16-week timeline, and the GPU tiers are sourced from the engagement archive; the shape of the marginal-cost curve is an illustrative amortisation profile whose endpoints are anchored on those sourced facts, and no price or per-document figure is asserted.

The horizon reader puts the four stages in the order they always arrive and draws the one line that carries the argument underneath them. Stand at the trapped-archive stage on the left and the marginal cost of the next document is high, because nothing has been amortised yet. Drag along the five-year path and the cost collapses toward a floor, because the fixed build spreads across more and more recovered documents while the training cost, paid once, does not recur. The four teal stage cards are the repeatable path; the single orange curve is what the path buys. The shape of that curve is an illustrative amortisation profile, flagged as such, but the direction is not illustrative and the endpoints are anchored on the sourced facts: a very large archive to divide across, and a fixed monthly GPU rent to divide.

Why "keeps falling" and not "is low"

It is worth being careful about the claim, because the sloppy version of it is wrong. The marginal cost is not low in some absolute sense on day one. On the first document recovered it is effectively the entire build cost, because there is nothing yet to share it with. The claim is about the trajectory: as the corpus is worked through, the per-document cost keeps falling, and it falls fastest early because the denominator grows fastest early. This is why the size of the archive is not just the problem statement from stage one, it is also the payoff in stage four. The same 136,771 files that made manual tracing impossible are what make the amortised cost per document small, because a large corpus is a large number of units to spread a fixed build across. The constraint and the reward are the same number seen from opposite ends of the path.

> **“The archive that made the manual route impossible is the same archive that makes the automated route cheap per document. The size is the problem in stage one and the payoff in stage four. It is one number read from two ends.”**

— FROM OUR OWN PORTFOLIO NOTES

VI

THE PORTFOLIO VIEW

What five years of the same path is worth

The repeatability is the asset

The reason to look at five years at once rather than one project at a time is that the repeatability is invisible from inside a single engagement and unmistakable across many. Any one project delivers a recovered archive to one operator. The portfolio delivers something the individual projects do not: the confidence that the next archive, in a curve set we have not seen, from a vintage of paper we have not scanned, will yield to the same four moves. Trapped corpus, synthetic bootstrap, learned segmenter, amortised recovery. The stages do not change. What changes between engagements is the generator's parameters and the operator's curve set, and both are parameter changes to a path we have already walked, not new paths.

That is the argument for treating the path as the asset. A model is a depreciating artefact; it is tied to a curve set, a scan vintage, a moment in the tooling. The path is not. The generator that manufactures training data is reusable. The segmenter architecture is reusable. The review-centred serving posture is reusable. The amortisation economics recur by construction. When we quote a 16-week accelerated build for a new engagement, the confidence behind that number does not come from the new engagement, which we have not started. It comes from having walked the identical path enough times that the stages, and the order they arrive in, are known quantities.

What actually varies between engagements

Honesty requires naming what is not constant, because the repeatable path is a claim about structure, not about effort being zero each time. The generator has to be retargeted to the new curve set, and if the new archive contains document types the generator has never modelled, that retargeting is real work. The scan quality varies, and a corpus of badly

degraded fourth-generation photocopies pushes harder on the model than a clean one. The cross-check coverage varies: where a rich set of already-digital LAS files exists, the recovered curves can be validated against ground truth, and where it does not, the review burden is heavier. The delivery timeline compresses or extends with the operator's deadline; 16 weeks was the accelerated track, and a standard track exists. None of this contradicts the repeatability. The four stages are the same; the amount of work inside each stage is what the engagement negotiates.

WHAT TO CARRY OUT OF THIS

1. Across five years the individual raster-to-vector projects converge on one repeatable path with four stages that always arrive in the same order: a trapped archive, a synthetic bootstrap, a learned segmenter, and amortised recovery. The repeatability, not any single model, is the asset.
2. The archive is the problem statement, not the input. On the reference engagement it was **136,771** scanned TIF files plus **7,781** LAS files, too large to hand-label and too varied to shortcut, with no supervised signal in it.
3. You cannot annotate your way out of an annotation problem, so the training set is manufactured. A synthetic bootstrap of **15,000** multiclass and **2,000** binary instances, randomised wide rather than made realistic, is what turns an impossible supervised task into a routine one.
4. Grade the segmenter on the recovered curve, not the mask. A peak curve-class IoU of **0.51** sits next to a peak R-squared of **0.9891** because a faithful curve survives imperfect pixels, and the review step puts a human on exactly the minority the model gets wrong.
5. Training is a one-off; recovery is not. A **16**-week build on a **750** to **1,800** EUR per month GPU is a fixed cost amortised across the whole corpus, so the marginal cost of the next document keeps falling, fastest early, as the archive it spreads across grows.

Limitations

This is a portfolio-altitude retrospective, and its numbers are drawn from one reference engagement to keep the argument concrete; the archive counts, the synthetic instance counts, the peak R-squared, the peak IoU, the 16-week accelerated timeline, and the GPU tiers are sourced from that engagement's records, and they should be read as one representative instance of the repeatable path rather than as a distribution across all engagements. The claim that the path repeats is a structural claim about the order and identity of the four stages, not a claim that the work inside each stage is constant; the generator retargeting, the scan-quality burden, the cross-check coverage, and the delivery timeline all vary, sometimes substantially. The marginal-cost curve in the instrument is an illustrative amortisation profile whose shape is presentational; only its direction and its endpoints, anchored on the sourced archive size and GPU rent, carry weight, and the instrument asserts no price and no per-document figure. The accuracy figures are peak values against a clean cross-check where already-digital LAS data existed; on portions of an archive with no ground-truth overlap the trustworthy-recovery claim rests on the human review step rather than on a measured agreement, and the honest limit remains that the segmenter is fast and accurate but not perfect. Finally, the acceleration context drawn from the subsurface literature describes the general value of digitising legacy records and the speedups digital inputs unlock downstream; it is context for why the path is worth walking, not a measurement of any one of our engagements.

References

1 Koroteev, D., Tekic, Z. (2021). *Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios for the future.* Energy and AI. The context for why decades of trapped subsurface records are worth recovering, and the source of the acceleration figures against manual and conventional workflows. https://www.sciencedirect.com/science/article/pii/S2666546820300033

2 Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P. (2017). *Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World.* IROS. The principle behind the synthetic bootstrap: randomise a generator's nuisance parameters widely enough and the real archive becomes just another sample of the synthetic distribution. https://arxiv.org/abs/1703.06907

3 Ronneberger, O., Fischer, P., Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation.* MICCAI. The encoder-decoder shape the learned segmenter sits in: it

consumes the scanned image and emits the per-pixel mask the vectoriser reads. https://arxiv.org/abs/1505.04597

4 Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., Dennison, D. (2015). *Hidden Technical Debt in Machine Learning Systems.* NeurIPS. The account of why the trained model is a small box inside a large system, which is why the falling marginal cost is a systems property, not a model property. https://papers.nips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

Get the full whitepaper

This page is the long-form summary. The complete whitepaper adds the stage-by-stage playbook for a new archive, the generator retargeting checklist for an unfamiliar curve set, the review-effort model that sets the human burden against the segmenter's accuracy, and the amortisation worksheet that turns a corpus size and a build cost into a per-document trajectory.

EARTHSCAN

Recovering Trapped Data: A Five-Year View of Industrial Digitisation

Published May 2025. Generated 2026-07-22.
https://www.earthscan.io/whitepapers/recovering-trapped-data-a-five-year-view-of-industrial-digitisation