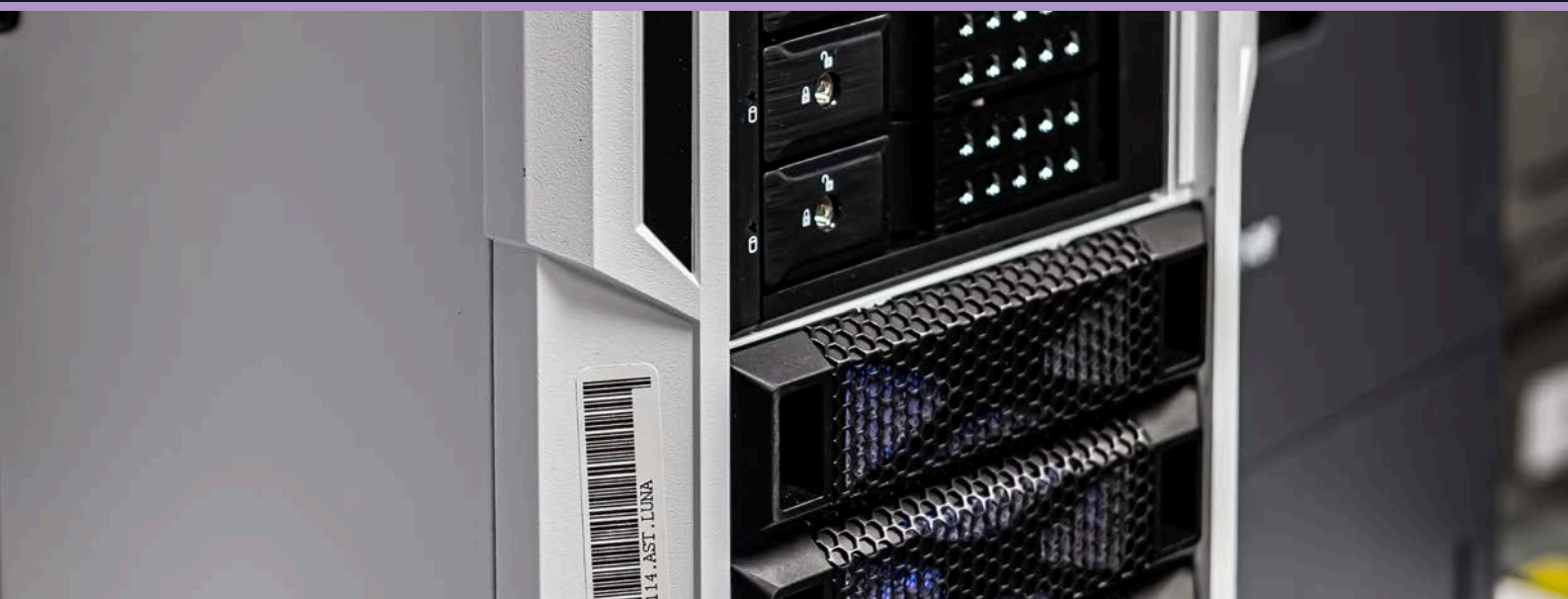
S2|DATA

The gamble of leaving data where it lies

Why the cheapest option is the one that costs the most

By Brendan Sullivan
CEO, S2|DATA



CONTENTS

The gamble

There is a decision being taken in many large organizations this year that nobody will document and nobody will sign. It is the decision to leave the data archive alone for another budget cycle.

It rarely feels like a decision. It feels like a sensible deferral. There is no live matter touching that data, no regulator asking about it, and no obvious return on the money it would take to find out what is actually in there. There is always something more pressing. So, the tapes or email archives stay where they are, the invoices from the storage providers, software and hardware support providers continue, and the question moves quietly to next year's list.

Say it out loud, though, and it is a bet. Leave the data where it lies, keep paying to store it, and hope that nobody asks a question you cannot answer.

> Most years, you win. That is exactly what makes the bet so easy to keep placing.

I have spent thirty years around this problem, and the organizations I meet are not careless. They are staffed by people who understand their obligations perfectly well and who have made a rational calculation with the information in front of them. Off-site storage is the stone you would rather not turn over, because you have no idea what is underneath it and every instinct tells you that finding out will be expensive.

What I want to emphasize here is why that calculation is usually wrong, what it costs on the occasions the bet does not come in, and why dealing with it now is a great deal more manageable than most people assume.

S2|DATA media vault

02 · DRIFT, NOT NEGLECT

Why the archive becomes a liability

The thing that surprises people is that the archive does not become a problem through neglect. It becomes a problem through drift, and the drift happens whether anyone touches the data or not.

Nobody did anything wrong

Picture a global company that created data in 2010. Its records managers did what good records managers do. They classified it: restricted, confidential, internal business communications, private, public. Anything sensitive, personal health information or personal identifiers, went in as restricted and quite possibly got encrypted. The retention schedules were set accordingly. Everything was documented. By the standards of 2010, this was a well-governed environment.

In May 2018, GDPR became enforceable, CCPA shortly after also bringing a different set of rules on what counts as personal data, what may be moved and what may not, and what an organization must do when that data is exposed or requested to be removed.

Here is where it gets awkward, and I want to be precise about it, because this is usually described far too loosely. It is not the case that everything classified in 2010 became wrong in 2018. Most of that estate was probably fine. But certain data could reasonably be classified one way before and another way after, and that is a very different problem from a clean break.

If every historical classification had been invalidated, you would at least know where you stood. You would have a defined piece of work in front of you. Instead, you have an estate in which some proportion of the classification still holds, some proportion no longer does, and action is required to comply.

Working out which is which after the fact is harder than it sounds. The person who classified that data in 2010 was making a judgment with context that may no longer exists. They knew which system it came from, what the business was doing at the time, who the parties were, and why a particular file was treated the way it was. That context is not known, relevant, or defensible. All that survives is the label.

This is worth being clear about, because it is where most conversations about legacy data start off on the wrong foot. Nobody made a mistake. The data was classified correctly, by competent people, with the best knowledge available at the time. The rules and situation changed underneath it.

> What that leaves behind is not a known gap that can be scoped and closed. It is doubt, spread thinly across the whole population, and doubt is considerably harder to answer for than a gap.

Then it compounds

Once you have that gap between the classification and the current rules, everything else that has happened over the last fifteen years makes it wider.

Obligations start to contradict each other. A US litigation hold requires you to preserve communications your own business policy would have deleted years ago, and which an EU data subject has a right to have removed. The same data is simultaneously must-keep and must-delete, depending on which jurisdiction is asking. Nobody drafted those rules with the other in mind, and you are left holding both.

The data itself is scattered across systems that were not designed to be read together. Some of it sits on network storage, some on tape, or media from other archives, and it matters which is which. A backup is a copy of something that exists elsewhere, taken so that the live environment can be rebuilt. An archive is a move, made deliberately, to hold material the business no longer needs to keep live. They were created for different reasons and they carry different obligations. Fifteen years later, they are frequently managed as though they were the same thing.

Then, from around 2016, everything moves to the cloud. New systems are stood up properly, with current rules, current classifications and current retention schedules. That part usually goes well. What goes with it is the reason the legacy estate now sits in a category of its own: it stays behind, governed by the old regime, while the organization's attention and its governance framework move on without it.

And the people go too. This is the part that does the most damage and gets discussed the least. The team that understood the original environment is often let go, or absorbed, or lost through an acquisition, precisely because the organization is moving to the cloud and no longer needs anyone to run the old stack. The obligation to keep the data survives. The ability and institutional knowledge to interrogate it does not.

Legacy is a condition, not an age

Which brings me to a definition I would ask you to sit with, because it changes who this article is about.

Legacy data is not old data. Data becomes legacy data the moment your organization can no longer access it, interrogate it, or explain it. That can happen thirty years after it was created. It can also happen in eighteen months, when a system is retired, a format falls out of support, a supplier contract ends, or the one person who understood how the thing was put together takes early retirement.

> The question is not how old your data is. The question is whether you could explain it if somebody asked you.

If you have been reading this thinking your organization does not really have a legacy data problem because you got off tape years ago, I would vigorously request testing that. The question is not how old your data is. The question is whether you could explain it if somebody asked you.

Not sure which side of that line you are on? Ask the S2|DATA team →

What it actually costs when it lands

The bet is rarely settled by choice. It is settled by an event, and there are three that tend to do it: a lawsuit that reaches back further than anyone expected, a regulator asking a question about a period nobody has thought about in a decade, or some form of a breach.

The legal and litigation exposure

When legacy data becomes relevant to a matter, the exposure is not really about the data. It is about your ability to produce it, and the courts have never been especially sympathetic to organizations that cannot.

The costs arrive in layers. There is the direct cost of the emergency exercise itself, which is always more expensive done under a deadline than it would have been done proactively. There is the risk of sanction where production fails or is incomplete. And there is the more damaging outcome: where an inability to produce is treated as something worse than an inability, and the court draws its own conclusions about what the missing material would have shown.

There is also a compounding effect that people underestimate until they are inside it. Data under legal preservation is a poison to the storage repository it is shared within. If there is one demand for preserved data then everything in that storage media must be saved also, running the risk of the same data becoming relevant and beyond the organizations ability to remediate through policy and retention schedules.

By the end of a long matter, an organization can find itself unable to dispose of anything at all. Every subsequent problem then lands on a larger pile.

The data and technical reality

Underneath the legal exposure sits a more stubborn problem, which is that the data may not be readable.

Tape degrades. Drives go out of warranty and out of production, and there is no guarantee an old one will spin up when you finally need it. Backup software was written to get data dumped quickly, spread across as many drives as the backup window time needed, but restoration of data was often less elegant. That is not a criticism of anyone in particular. It is simply how the technology developed, dictated by data volume and backup window allotment, and the consequences land twenty years later.

Then there are the applications used by the organization. Data written by a system that no longer exists is not self-explaining. A database in an obsolete format is not a database any more, it is a file. Restoring the bytes is one problem. Understanding what they meant is another, joining critical data parts together a challenge, and it is the harder of the two.

This is why the honest answer to “can we get at it?” is nearly always yes, and why that answer is less reassuring than it sounds. The question worth asking is what it costs, how long it takes, and whether you find out on your own schedule or somebody else’s.

The information governance problem

The third angle is the one that gets least attention and, in my experience, causes the most sustained discomfort.

If an auditor or a regulator asks you to demonstrate that your data is classified correctly, you can almost certainly do it for everything created in the last few years. The cloud environment was built for today's data governance requirements. The classifications are current, the retention schedules are applied, and the reporting is there.

> Then they ask about 2010.

At that point the honest answer is that you do not know, and the reason you do not know is everything described above. That gap is not a technical embarrassment. It is a governance failure that sits on your whole estate, because a governance framework that covers the easy eighty percent and cannot speak for the rest is not really a framework at all.

A data breach is where this becomes most acute, and it is the reason this article is as relevant to a CISO as it is to a General Counsel. Regulatory notification obligations require you to say, quickly and accurately, what was affected and whose data it was. If the compromised material sits in an archive nobody has been able to characterize for a decade, you cannot answer that question inside the notification window. You are then choosing between over-notifying, which carries its own cost and reputational damage, and under-notifying, which can be considerably worse.

Do you know what is sitting in your archive?

If the honest answer is not yet, that is the normal position. Establishing it is quicker and less expensive than a restoration.

Talk to the S2|DATA team →

05 · METADATA, NOT RESTORATION

It does not have to be this way

Here is the part that changes the calculation, and it is the reason I think most organizations are getting the math wrong rather than being reckless.

The assumption behind leaving the archive alone is that dealing with it means restoring it. Landing all that data somewhere, processing it, reviewing it, and paying for the privilege at a scale that makes the whole thing a non-starter. If that were the only route, the deferral would be entirely rational.

> In most cases you do not need to touch the actual data at all. You need to analyze the metadata.

The ratio is what makes this work. An organization might hold fifty terabytes across a legacy estate and only a few hundred gigabytes of metadata describing it and identifying data's original location, owner etc. That metadata often tells you when data was created, who it involved, what the backup sessions were, what the subject lines said, and which systems it came off. That is frequently enough to make sound, documented business decisions about material you have never opened.

An example of what that buys you. Suppose an organization ran ten Exchange servers in 2010. One custodian sat on server three, another on server four, and the first is subject to a hold while the second is not. If the metadata tells you which server each mailbox was backed up from, you can remediate the second custodian's data by dealing with server four, without going anywhere near the material you are obliged to preserve. You have separated the two populations without restoring either.

Multiply that across an estate and the position changes completely. Instead of holding everything indefinitely because you cannot tell the difference between what must be kept and what could go, you start disposing of data you were always entitled to dispose of, on your own timetable, before the next hold lands and freezes it in place.

This is what the S2|DATA TRACS engine was built to do: scan media, extract metadata at scale, and produce a structured index of an estate without landing the content. Combined with an early data assessment, it gives you a defensible picture of what you are holding, at a cost that bears no relation to a full restoration. And because the underlying media is left unrestored, you keep every option open. Nothing about the exercise commits you to a course of action.

None of this removes the need for judgment. Metadata tells you what you have and where it sits, not what it means or what you should do about it. That still requires people who understand both the technology and the obligations. But it does mean the first step is small, quick and low cost enough that there is no longer a good reason to keep deferring it.

Two ways in

In practice, organizations come to this from one of two directions, and both are legitimate.

The first is preparation. There is no live matter, but there is a recognized gap, and somebody has decided they would rather know than not. This is by far the better position to be in. Work done calmly costs a fraction of work done under a production deadline, and it is the only version of this exercise where you get to make decisions rather than react to them.

The second is reaction. Something has already happened. There is a matter, a request, or an incident, and it reaches into data nobody can currently account for. The work is the same, but the clock is running and the options are narrower.

The calculation that decides which of these you are having is simple enough, and I have watched it play out many times. If the exposure is significant and the fix costs more than the exposure, no rational board approves it. They take the hit. But when the fix is a fraction of the exposure, and particularly when it also prevents the next one, the decision usually makes itself.

That is the point worth holding on to. The reason most organizations do not act is not necessarily that they have weighed the cost and decided against it. It is that they have assumed a cost far higher than the real one, and never got as far as the weighing.

Preparing, or already in the middle of it? Scope it in one conversation →

What good looks like

The destination here is not an empty archive. It is a straightforward one: being able to show that your classifications are current and relevant across the entire estate, historical and current, at any point somebody asks.

That means being able to demonstrate not only that data is classified correctly today, but that you have retained the ability to act on it historically, whether that means deleting, remediating, migrating or restoring. It means no gap in the record between the systems you run now and the ones you retired years ago.

It is a shift in posture, and I think it is the most useful way to understand what has changed. For twenty years the discipline tried to control data at the point of creation, classifying and categorizing everything as it arrived, in the hope of staying on top of it. That was always going to be a losing battle: the number of data sources has accelerated every year and shows no sign of slowing. What is now possible is different. Rather than control everything at the start, you can understand it at the end, when a question is finally asked and you know what you are actually looking for.

Defensibility is what makes that stand up. Not as a slogan, but as the ordinary discipline of documented process, chain of custody and a reproducible audit trail behind every data decision. It is what allows you to say to a regulator that you did not simply delete a population of data, you established what it was, applied a rule, recorded the basis, and can show your working.

None of that is exotic. It is the difference between an organization that has an answer and one that is hoping the question does not come.

Next step

If you are not sure what is sitting in your archive, that is the normal position, not an unusual one. The useful first move is a short conversation about where your exposure actually sits and what it would take to establish a clear picture of it. Most of the time, that first step is smaller than people expect.

Find out what you are holding before somebody asks

Book a short scoping conversation with the S2|DATA team. We will establish where your exposure sits, what an assessment would cover, and what it would cost.

Talk to the S2|DATA team

s2data.com · info@s2data.com

Frequently asked questions

What if all of our data is stored on backup tapes?

Tape is one of the most common legacy formats and one of the most workable, provided it is handled properly. Metadata can be extracted from tape at scale without restoring the contents, which gives you a searchable index of what is on each tape and lets you identify the material that matters before committing to any restoration.

What if the application that created our data no longer exists?

This is routine rather than exceptional. Data written by a retired application can still be read at session and file level, and metadata extracted, without reinstating the original system.

What if we cannot access our archived data any longer?

Obsolete media, unsupported formats and missing hardware are solvable with the right equipment and expertise. The more useful question is what recovering it would take, and whether you establish that now or under the pressure of a deadline.

What if the only person who understood this system has retired?

Lost institutional knowledge is one of the most common reasons data becomes legacy data, and it does not have to be recoverable for the data to be. Metadata extraction rebuilds a structural picture of an environment, including how it was organized and where material resided, without depending on anyone remembering how it was built. The list of projects we have tackled over thirty years makes S2|DATA the “keepers of lost knowledge”

What if we receive a lawsuit and cannot access the data?

An inability to produce is treated far more seriously than an inability to find quickly, and courts may draw adverse conclusions from missing material. Legacy data can be assessed and produced under deadline, but it costs considerably more than the same work done in advance. If a matter is live, the priority is establishing scope quickly.

What happens if regulators ask for information we cannot produce?

Regulators generally expect an organization to account for its whole data estate, not only the systems currently in use. Being unable to answer for a historical period is treated as a governance failure rather than a technical one. The practical answer is to establish what exists before you are asked, so the response is a matter of retrieval rather than discovery.

What if our backups are unreadable?

Backups that fail to restore through their original software can frequently still be read at a lower level, because the data is usually intact even where the catalog or index is not. Physical damage to a backup is a genuine risk that increases with time, which is the strongest argument for assessing an aging estate sooner rather than later.

How do I know whether my archived data is still usable?

The only reliable answer comes from testing a representative sample rather than assuming. A structured assessment establishes what can be read, what has degraded, and what is at risk, and produces a documented picture of the estate. That assessment is inexpensive relative to a full restoration and is the sensible starting point.

ABOUT THE AUTHOR

Brendan Sullivan

CEO, S2|DATA

Brendan has spent more than three decades in data storage, from the early engineering of enterprise tape cartridges to leading global recovery projects. He founded S2|DATA to solve the problems that sit deepest in legacy systems and forgotten archives, and has since supported law firms and corporate legal departments through matters involving regulatory pressure, unexpected data loss and questions that fall outside the reach of standard tools.

His view, and the argument of this paper, is that everything becomes legacy data eventually. What matters is whether an organization can still explain it when someone finally asks.

Connect with Brendan: LinkedIn

About S2|DATA

S2|DATA restores, migrates and makes sense of legacy data: the archives, backup tapes and retired systems an organization can no longer access, interrogate or explain.

We work with corporate legal departments, law firms, compliance and information governance teams on the part of the estate that the large governance programs never reach. We recover data others have written off, extract metadata without the cost and risk of a full restoration, and move ageing email archives onto a single searchable platform, so that decisions about what to keep, remediate or produce can be made on evidence rather than assumption.

Our TRACS engine scans and extracts metadata as media is read, without landing the content. Invenire replaces end-of-life archive systems such as Dell EMC SourceOne and Veritas Enterprise Vault with one cloud platform, precise discovery and on-demand export, for a single monthly fee.

OFFICES

Atlanta | Austin | London

CONTACT

s2data.com
info@s2data.com
+1 678 626 1659

Turning Complexity Into Capability™

S2|DATA

Do you know what is sitting in your archive?

Most organizations cannot answer that with confidence, and finding out is quicker and less expensive than it tends to sound. A short conversation with the S2|DATA team will establish where your exposure actually sits.

Talk to the S2|DATA team

s2data.com

Turning Complexity Into Capability™