ZL UNIFIED ARCHIVE®

What is Dark Data? The Risks of ROT

ZL TECHNOLOGIES | White Paper

ZLTECH

Much of the information stored in file shares amounts to “dark data” – content that is accumulated by business users but left relatively unmanaged and un-monetized.

The Gartner definition of dark data is as follows:

> “**Dark data** is the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes (for example, analytics, business relationships and direct monetizing). Similar to dark matter in physics, dark data often comprises most organizations’ universe of information assets.”

Given only superficial assessment, dark data does not seem to be a critical business concern; it is a byproduct of an active and productive business ecosystem. But just as industrial manufacturing byproducts may be toxic at scale, so too may the digital products of a large organization. Because dark data takes up such a large portion of enterprise content, the digital debris and buildup contained within file shares can cause numerous problems that erode business profitability over time.

The most taxing subcategory within dark data is frequently deemed “ROT” – Redundant, Obsolete, and Trivial content. These are the unfinished word documents from employees that left long ago, the images that workers downloaded to update various social platform profiles, and the saved videos of the 2010 holiday party karaoke contest. For the most part, ROT data largely exists in file shares.

Even though ROT data only represents a subset of ROT content and is not inherently bad on its own, ROT data can still cause several major business problems.

(1) Lost Productivity

(2) Legal and Regulatory Risk

(3) Excess Storage Demand

ROT -- Lost Productivity

Lost productivity, while incremental and often not noticeable at the individual level, can add up to crippling losses within a large organization. File shares are often a key culprit, as employees tend to hoard and duplicate content that may only be useful for a short period of time or may be completely irrelevant to business tasks altogether. This tendency, combined with the human habit to inconsistently name and organize data within the file share, makes for an environment that is extremely messy and difficult to search. A well-intended search for a given document may take a significant amount of time, or may not even yield the correct (or most up-to-date) item.

Lost productivity, even little bits of inefficiency, add up to significant costs for large organizations. For an example, assume the following conservative estimates:

- 10,000 business users in a firm
- 15 minutes per day spent on unsuccessful file share searches
- $15 per hour wages
- 48 full 5-day work weeks in a year

Given the above parameters, the firm stands to lose $9,000,000 over the course of a year simply from unproductive file share searches. While there are many, many other sources of impaired productivity, file shares stand out because they are (1) an infrastructural issue rather than a behavioral issue, and (2) because they are nearly universal in their use as the vast majority of employees access them multiple times per day. In theory, training programs could be implemented to help individuals better manage their own files; however, such an approach would likely be costly, time-consuming, and ultimately ineffective. The root of the problem is the structure of file shares themselves and the nature of the content they collect.

ROT -- Legal and Regulatory Risk

Excess and outdated content can also create problems for compliance and legal teams. Traditionally, this issue has been dealt with via retention policies and consistent rules for the lifecycles of documents. However, no matter what technology tool is used to execute these policies, defensible deletion of content is impossible when copies reside in multiple locations scattered all throughout the enterprise. Enterprise content management solutions, which are commonly leveraged to manage the lifecycle of items that have been identified as relevant to business, only manage data that is actively placed into the platform – usually by records managers. The problem with this approach is that it leaves the vast file share environment untouched, where duplicates may remain indefinitely.

Lifecycle control for formal “records” alone is not enough, since the Federal Rules of Civil Procedure (FRCP) state that practically any data within the business is discoverable during a lawsuit. Content outside of ECMs and other repositories – such as email archives – are still fair game for discovery requests, as long as there is a reasonable chance that the data will have relevance to the case. The old adage that states, “if you’re not doing anything wrong, you have nothing to worry about” doesn’t quite hold true in the context of litigation and eDiscovery, which is often interpretive in nature. So seemingly innocuous outdated or trivial content may become relevant in unexpected and detrimental ways.

The best defense against unpleasant surprises in the litigation process is the establishment of clearly-defined, consistent policies for content. In most cases outside of certain regulations which require exact retention requirements, courts care most about consistency. If emails are kept only for 6 months before permanent deletion, it is likely defensible if – and only if – that policy was already well-established before the case and was executed with precise consistency. The problem with this, of course, is that content needs to be well-managed for policies to be executed without variance. And file shares are often the biggest offender for unmanaged, human-created content.

ROT -- Excess Storage Demand

Storage cost is not as much of a business concern as it used to be; prices have drastically fallen over time for hardware, and the rise of private and public cloud storage options have given businesses more flexible options for the offloading of content. However, in large organizations, the cumulative cost of storage can still be a major concern – especially when using collaboration platforms such as Microsoft SharePoint which require additional servers to scale up, and have few or no options for lower-tier storage.

File shares, in particular, can build up content quickly and exponentially. File environments have the inherent tendency to spawn duplicate data copies; a mass email sent out with an individual attachment may be saved individually by 100+ recipients within the same file share, even without any modifications or edits. For simple text documents, this burden can certainly mount over time. But with even larger files, such as presentation decks and graphic design files, storage requirements can increase significantly. Business users rarely take the time to dispose of outdated, duplicated, or irrelevant content themselves, and without an underlying governance platform, this issue of duplicated content is both exponential and unresolvable.

It's clear that dark data poses not only an IT burden, but an enterprise-wide burden. Information grows and becomes duplicated and scattered as individual users classify document copies as they see fit. The answer isn't to re-shape human behavior; it's to create a governance system that governs content without impeding the day-to-day workflow of business users.

This isn't to say that ROT is purely without value; in fact, in today's big data era, some might argue that even outdated and trivial content has value for analysis by showing changes in business practices and worker habits over time. On the other hand, many others still firmly believe that ROT's marginal value pales in comparison to the attendant risk that accumulates with messy content. But whichever side of the fence your particular organization falls on, one thing is clear: ROT data is impeding the efficiency of file share environments. Even if ROT is to be kept and stored for future analysis, the file share is not the ideal place to do so.