

Business Content Landscape Report

Your Guide to Unstructured Data Analysis

This report provides an overview of how creating an inventory of your unstructured data can help identify redundant, obsolete, and trivial (ROT) data, so you can eliminate it and drive productivity.

The Problem of Exponentially Growing Data

In a world with no shortage of data storage solutions, a significant problem is being hidden in plain sight—the exponential growth of data. It’s estimated that by 2025 global data creation will grow to more than 180 zettabytes (one zettabyte equals a trillion gigabytes).¹ This is an astounding amount of data, and only a small fraction of it is structured. Structured data refers to numbers and data fields that can be organized, filtered, sorted, and analyzed. In other words, only a small percentage of enterprise data is ready for business use.

Experts estimate that unstructured content already accounts for up to a staggering 80 to 90 percent of all digital data,² much of which is locked up in various file formats, and stored across many systems and repositories. The vast majority of data being created is both difficult to manage and difficult to extract value from. This may seem like an insurmountable problem, but this high volume of (unstructured and structured) data is a huge, untapped resource with the potential to create a competitive advantage for companies who know how to use it.

To tap into this competitive advantage, organizations need total visibility and control of their growing data to reveal hidden insights. This process begins with building a comprehensive inventory of unstructured data and its existing metadata from all sources, leading to a better understanding of what you have, where it is, and how much there is.

¹ IDC: Expect 175 zettabytes of data worldwide by 2025, *Network World*

² Tapping the power of unstructured data, *MIT Management Sloan School*



UNSTRUCTURED DATA:

Information that either does not have a pre-defined data model or is not organized in a pre-defined manner (text, video, audio, web server logs, social media, etc). It usually resides in file shares, Content Management Systems (CMS), Document Management Systems (DMS), etc. Minimal metadata exists on each file, and the metadata that does exist is highly dependent on the classification system, originating application, and what people entered while creating it. Unlike structured data, unstructured data has ample information and value, but fewer identifying characteristics to make it easy to find and use.



STRUCTURED DATA:

Typically categorized as quantitative data. It is highly organized and easily decipherable, especially by machines. It lives in databases and systems built on top of database structures. Many tools exist that can ingest, analyze, and visualize unstructured data.

A comprehensive inventory enables efficiencies, beginning with cleaning up ROT

To prepare for data-centric initiatives, whether it's a broad, enterprise-scale digital transformation, tactical process improvements, or risk reduction projects, you need to gain an in-depth understanding of your data.

The first step is to build an inventory that identifies all of your organization's files, documents, and data to create a complete picture of your data ecosystem. This is what Shinydocs refers to as your Content Landscape.

Assessing your Content Landscape allows you to pinpoint the documents and files with real business value and manage those that no longer have value through the end of their lifecycle, dealing appropriately with your redundant, obsolete, and trivial (ROT) information.

Reducing ROT provides efficiencies in data and information lifecycle management, such as reducing storage costs. It also reduces legal, business, and regulatory compliance risks.

Let's walk through our recommendations for managing ROT using the average crawl data presented below as a case study.

Average files examined in one day

Average files crawled per second: 863

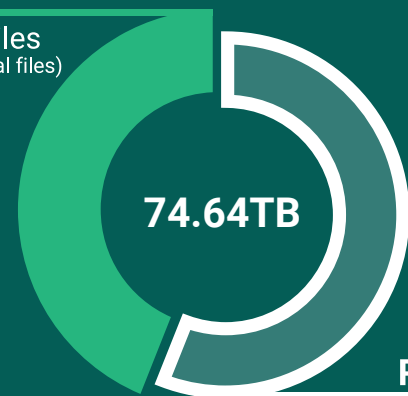
Average files crawled per day: 74,638,080

Average TB crawled per day: 74.64 TB

(Example scenario based on the amount of data we can crawl at the current speed)

Files of Value

34.3M files
(46% of total files)



ROT

(Redundant, obsolete, and trivial files)

40.3M files
(54% of total files)

\$136,617 Savings @\$3.39/GB

Redundant Data

DEFINITION:

Redundant data occurs when the same piece of data exists in multiple places in the same system or across multiple systems. Redundancy can cause data inconsistency and lead to issues with different versions of a file or document, which can provide a company with unreliable and/or meaningless information, when employees do not know which version they should be using.

It is difficult to create a single point of truth (SPOT) for data when you do not have a clear understanding of how much redundant data there is and where it resides.

OUR APPROACH:

Each organization may have its own view of what constitutes redundant data. We work closely with every organization to determine the best set of redundant, obsolete, and trivial (ROT) rules to follow. We know that context matters and rules must be established for an organization's specific data environment.

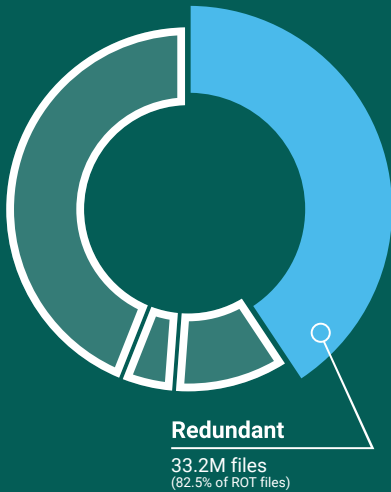
NEXT STEPS:

After building your inventory to assess your Content Landscape, Shinydocs can work with you to define redundant data in your context, for example:

- Simple comparison of file names and other metadata of files from different sources
- De-duplication using a computer algorithm known as a 'hash' which produces a unique identifier for every binary file, and comparing these to produce a highly accurate report of duplicate files

OUTCOMES:

To improve the efficiency of many processes, it is important to remove unnecessary redundant files. This frees up space in primary storage and in backups, speeding up backup and subsequent recovery processes. This results in reducing costs related to storage, disaster recovery, and business continuity. It also reduces the business risk of employees using the wrong version of a file or document.



REDUNDANT RULES

- Traditional email file extensions: .boe, .box, .eml, .mbox, .mbs, .mbx, .mmdf, .msf, .msg, .nsf, .ost, .pst, .tbb or folder named "email archive"
- Software images with extension of .iso, .img with a file size larger than 2GB
- File extension of .class, .jad, .jar, .java, .jp, .idx, .cod, .j

Total Number of Redundant Files:

33,247,500 files

Obsolete Data

DEFINITION:

Obsolete means “no longer in general use”, “discarded”, “replaced”, or “out of date.” Obsolete data includes content that is incorrect, has been superseded or replaced, or is simply no longer in use. It is a broader category than content which is simply outdated.

If not disposed of, outdated, obsolete, and forgotten files can remain unknown and on servers forever. These files may contain sensitive information, exposing your organization to business, legal, regulatory, and reputational risk.

For instance, the 2014 Sony hack revealed embarrassing email exchanges that under Sony policy should have been deleted, exposing executive salaries, gender disparity in wages, casual racism, and unflattering messages that did immense damage to the brand.

OUR APPROACH:

We work with you to determine your criteria for assessing obsolete data, and files containing out-of-date information by analyzing created, modified, and accessed date related metadata.

NEXT STEPS:

After building your inventory to assess your Content Landscape, Shinydocs can work with you to reduce the number of obsolete files:

- Reduce file count by disposing of files based on easily applied business rules; for example, files older than 5 years that have not been accessed in the last year.
- Archive. Organizations nervous about defensible disposition can instead opt to move these files to inexpensive cold storage but still ensure the information is searchable across the organization.
- Create more comprehensive rules incorporating obsolete and trivial files. e.g. uncovering .log files that are older than 1 year.

OUTCOMES:

Ensure employees are not using out-of-date information within their business processes.

Reduce the risks involved with retaining obsolete and redundant information. This includes compliance with new privacy regulations being implemented across the globe regarding personal information. Such privacy laws state how long an individual’s information can be kept.



OBSOLETE RULES

- File extension .log with create date older than 6 months
- Files/folders including the word “Draft” with a last modified date older than a year
- Last modified date older than 7 years

Total Number of Obsolete Files:

4,916,600 files

Trivial Data

DEFINITION:

Trivial data is information that does not need to be kept. Specifically, trivial information is content that does not contribute to corporate knowledge, business insight, or record-keeping requirements.

Trivial files like .tmp and .log files simply take up storage space and can muddy up enterprise search results, returning files that are irrelevant.

Working with a number of customers, we've uncovered terabytes of trivial information by extending search parameters to include .mp3 music files, extensive .avi or .mkv movie downloads and even entire .iso gaming libraries.

OUR APPROACH:

We work with you to identify trivial files by type as well as in combination with age.

NEXT STEPS:

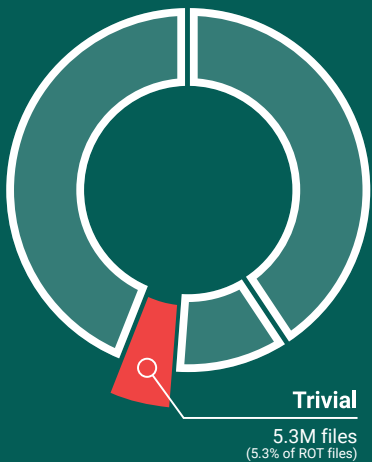
After building your inventory to assess your Content Landscape, Shinydocs can work with you to develop a tailored strategy for:

- Finding trivial data
- Archiving for further review if required
- Supporting deletion of trivial data

OUTCOMES:

Outcomes are very similar to those realized by removing redundant and obsolete files, including freeing up space in primary storage, and in backups, speeding up backup and subsequent recovery processes.

Removing trivial files is a simple way to reduce the size of your overall Content Landscape, improving efficiency for business and IT operations.



TRIVIAL RULES

- File extensions .\$\$\$,.old, .bak with a last modified date older than 30 days
- File name thumbs.db
- File size of 0 bytes
- File extension = <null>
- File name contains ~ character
- File extension: .tmp, .temp
- Folders containing any of the following terms: torrent, magnetlink, pirate bay

Total Number of Trivial Files:

2,135,900 files

We help you find, understand, and take actions on all your data.

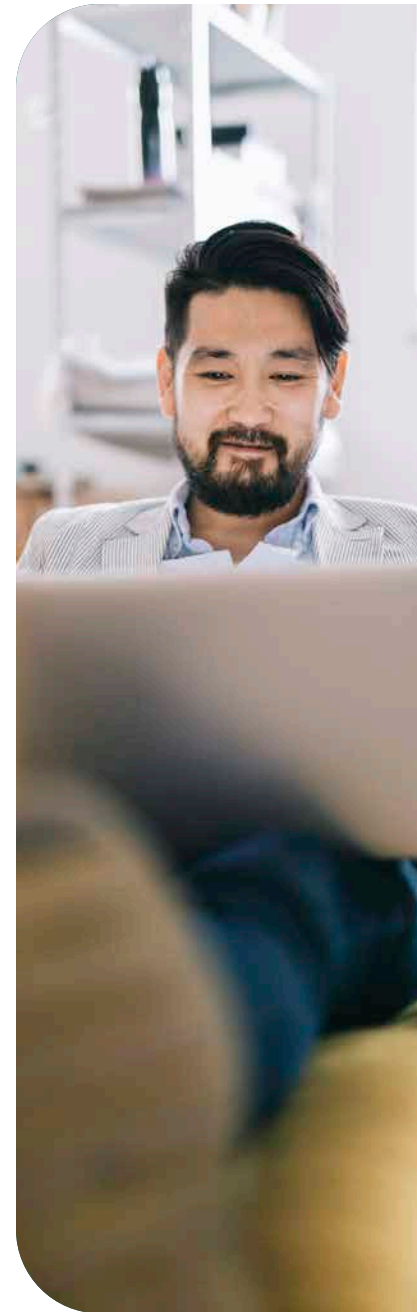
Shinydocs software creates an inventory of the content stored across your organization and enables you to quickly find information, automatically govern files, migrate efficiently, and maximize value as your data grows.

When a comprehensive content inventory from across enterprise repositories has been created, it:

- Provides insight into all your structured and unstructured data, content, and information.
- Enables you to eliminate redundant, obsolete, and trivial files, shrinking your digital footprint.
- Delivers a modern search experience across all enterprise information.

Shinydocs products simplify the process of finding, understanding, and getting value from your data:






- **Data Discovery:** Find and monitor content across your organization with our data discovery platform and get insights to improve your business and manage risk.
- **Enterprise Search:** Find the answers and insights your teams need by accessing any file across your organization with a simple and powerful enterprise search.
- **Data Migration:** Move your files, documents, records and media sooner to get value from new systems faster and store content in the right locations.



About Shinydocs

Founded in 2013, Shinydocs' software automates the process of finding and identifying all files, media content, and documents buried in repositories across your organization so you can make more informed decisions to drive growth and positive customer experiences. We believe that there's a better, more intuitive, and cost-effective way for organizations to manage their information. This belief has propelled us forward in the development of our software, services, and strategy and allows us to support global corporations in their pursuit of improving their customer experience, collaboration, innovation, and ability to demonstrate compliance with industry regulations through easier and more effective management of their business content.

For more information visit: www.shinydocs.com

-  @Shinydocs
-  @Shinydocs
-  @Shinydocs
-  [linkedin.com/company/Shinydocs](https://www.linkedin.com/company/Shinydocs)
-  info@Shinydocs.com