CONNEXUS

INFORMATION GOVERNANCE

TECHNICAL WHITE PAPER

v1.0: 11/16

CONTENTS

EXECUTIVE SUMMARY

Why does Information Governance matter? Just a few decades ago, filing clerks would meticulously sort and organise paper records within physical cabinets, ensuring that correct records were labelled and stored in the correct location, removing outdated and duplicate copies. Now that users file electronic documents themselves on apparently cheap and invisible storage, high value business information which supports your ongoing success is frequently buried and often lost within chaotic unmanaged repositories.

A common but significant risk is that the surplus information is often regarded as just irrelevant e-Trash or Redundant, Obsolete and Trivial (ROT), but in addition to making the right information harder to find, you are likely to be storing potentially dangerous content, which could put your organisation at significant risk. This content should be properly managed or remediated for risk mitigation.

Giving relevant staff straightforward access to the right information is a cornerstone of business success, so it is the organisations which embrace information governance that are giving themselves the best chance to excel.

When implemented well, information governance can help you quantify and determine the true cost of each category of information, giving you a clear and actionable policy and methodology to plan and deliver an ongoing course of action to take control of information within your organisation at minimal cost.

BENEFITS OF INFORMATION GOVERNANCE

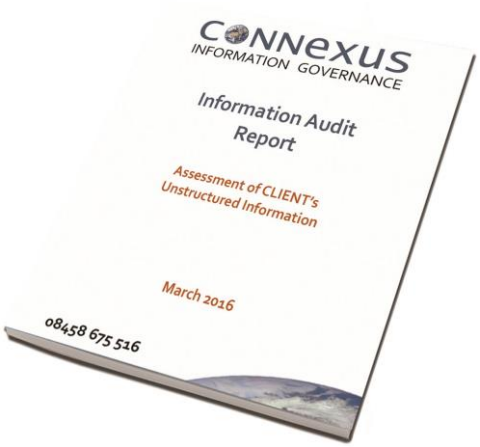
- Storage savings - clarity on what to remove and what this can save you
- Increased user productivity - find the right information, previously buried in a mountain of data
- Enhanced user collaboration - easier access to shared documents
- Retention policy applied - remove files when they become obsolete
- Sensitive data secured - ensure suitable management of restricted files, reducing the risk of accidental or even malicious exposure, such as via cyber-attack.
- Data loss prevention - ensure protectively marked and sensitive data has not leaked into areas where it is at increased risk of Cybersecurity breaches
- Migration much easier - migrate a lower volume of high value files with consistent metadata

CONNEXUS IG SERVICES

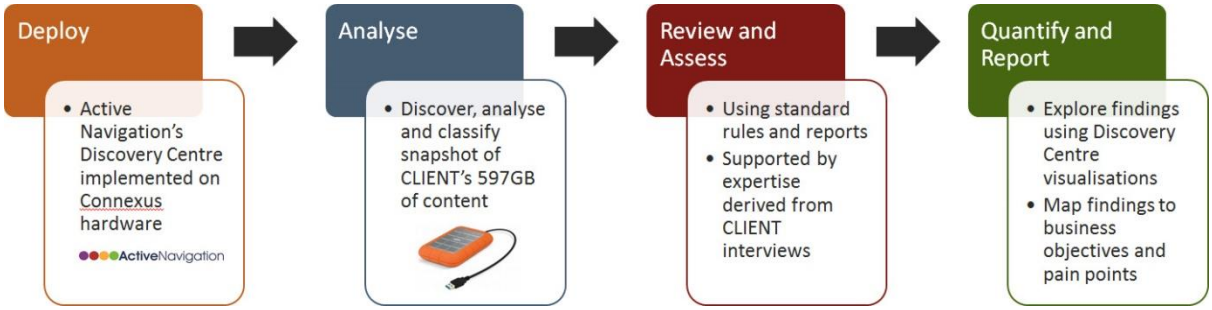
If you're swamped by a mountain of irrelevant data, can't be sure that you're working with up to date information and are worried what dangers might be lurking in your files, Connexus IG offers the World-class consulting and revolutionary software to make it easier for you to succeed.

INFORMATION AUDIT

The Connexus IG software-assisted Information Audit is commonly used to help build a business case for information governance, locating and quantifying different categories of content (we have many default categories and the service includes expert services to find your custom records), calculating their value and prioritising a plan of action to remove, quarantine, protect, tag and / or migrate your files.



The service clearly identifies surplus and at risk files within the legal, regulatory, and business requirements which apply to your organisation. The Information Audit determines the relative cost and priority of remediating each category of information, which can form the basis of an information governance strategy. The detailed information contained in the final report will help you prepare and justify your business case for information governance.



The audit process is the springboard from which all other aspects are launched because it removes the guesswork from planning and places facts at the centre of the discussion. Just as importantly, at a time when business functions and information owners understand the principles behind good governance but struggle with justification, the audit makes it clear where issues lie and the impact they are having. In turn, audit results provide a strong lever for changing user behaviours, enabling them to see the effects of their information practices and culture – it is hard to deny poor collaboration and sharing when large swathes of content can be shown to be in personal drives or laptop hard drives. As an educational tool, audit results have great impact.

- Identification and correction of poor information management and usage habits and cultures (such as the use of personal drives at the expense or in duplication of shared collaborative environments).
- Fact-based evidence of content distribution and trends drives the creation of content management policies, enhancement of acceptable use policies and in support of user education and training.

Complemented with an interaction with information owners and users, content inventories are the foundation for good policy, allowing policies to be developed based upon real data rather than informed supposition. Typical audit results allow Acceptable Use policies to be tested and enhanced and establish a need for policies and action for content clean-up.

POLICY GUIDANCE

A logical set of clear rules captured in a set of policy documents are an essential foundation for any information governance project. Policies such as Acceptable Use, Data Protection, Records Management and Dispoal are the essential link between legislated corporate responsibilities and business user compliance. Information governance policies should comply with the relevant external regulatory requirements and directives, which are essential as the foundation on which a repeatable and defensible process is built. It is these clear guidelines which enable you to identify and deal with violations in order to ensure that policy is adhered to. A review of your existing policy, particularly when combined with the quantitative results of an Information Audit will help identify any weaknesses and identify areas for improvement. Connexus IG will review your policies and highlight deficiencies and enhancements in line with relevant applicable legislation.

PROJECT PLANNING

A well defined project plan with measurable outputs will of course help to ensure progress is being made as intended. Understanding the optimum order of tasks, effort required and stakeholder responsibilities is key to bringing this together into a coherent plan. For example, it is essential to gain support from your organisation's executive management, as this will enable business user engagement for the project to progress as intended. Business users often struggle to justify contributing to information governance projects without clear directives from those leading the business. Encouraging the business to adopt your information governance processes one of the key challenges you will need to overcome, but we have many tips and tricks to make this more effective. It may sound obvious, but we have seen too many examples of information governance ambassadors trying to engage single-handedly with the whole business. You will also need to build a supportive network of stakeholders throughout the business for the best chance of success. As a minimum, key representation from each business area should contribute department specific requirements and provide delegated access to experienced members of staff for relevant approvals. Involving IT is often critical, as gaining a rapid and deep understanding of information is usually only viable with software support.

It sounds too simple to need to state, but we have seen many organisations attempting to implement steps to improve information governance centrally without considering that the whole business is affected. Everyone will need guidance on how to improve the situation, otherwise centralised efforts to cleanse and reorganise data will leave users confused, frustrated and fighting to restore to the original problematic state of affairs.

Do take the time to share your success stories, as a good case study is surprisingly motivating for more reticent heads of departments. We have seen cases where some departments need to be reassured by seeing a plan working effectively before they consider adopting it, and other cases where simple departmental rivalry has led to a competitive approach to achieving the best results.

Our experienced consultants will bring all of these elements together, offering clear advice on the most effective means to engage for project success.

ROT REMOVAL

Redundant, obsolete and trivial content which gets in the way of business content with real value should be deleted (or quarantined to cheaper secondary storage) to improve user productivity and reduce storage costs. When factoring in all costs associated with primary enterprise storage, Gartner estimates that the annual cost per TeraByte is up to £15000, which can sometimes be saved directly, but more frequently result in a postponed procurement to extend current storage capacity. We have developed definitions of up to 50 types of ROT which can be implemented quickly to gain rapid momentum in the implementation of your project. As tasks are policy-driven and fully audited, the whole process is defensible.

Redundant, obsolete and trivial content which gets in the way of business content with real value should be deleted (or quarantined to cheaper secondary storage) to improve user productivity and reduce storage costs.

When factoring in all costs associated with primary enterprise storage, the annual cost per Terabyte is likely to be up to £15,000, which can sometimes be saved directly, but more frequently result in a postponed procurement to extend current storage capacity.

Benefits of removal of ROT are:

- Removal from storage, reducing the load on managed storage infrastructure and simplifying user information retrieval.
- Reduced load and cost for content migration or preparation for legal cases in the event of litigation.

Connexus distinguishes between ROT which can safely be removed (high confidence), that which will require a small amount of batch sampling (medium confidence) and that which is likely to require a more time consuming file by file review (low confidence), so you can prioritise the subsequent deletion or quarantine actions depending on levels of confidence and business user time available for review.

ROT - HIGH CONFIDENCE

Documents containing prohibited content is proposed for deletion, as the level of confidence identifying it as ROT with no business value is high, with low risk anticipated in impact on the business. After review by a central Information Governance team, it is proposed that these rules be signed off centrally and implemented rapidly without further business user review. If files are deleted in error, IT backups are available for restore, but the risk of this being required is very low. The following default categories of high confidence ROT rules are fully defined and tested:

- REDUNDANT RECYCLE BIN
- REDUNDANT TEMP & BACKUP
- OBSOLETE LOG FILE
- OBSOLETE NAMED BACKUP

- OBSOLETE NAMED COPY
- OBSOLETE NAMED DELETE
- OBSOLETE NAMED OLD
- OBSOLETE NAMED SUPERCEDED
- TRIVIAL NAMED FILM
- TRIVIAL NAMED MUSIC
- TRIVIAL NAMED TV

ROT - MEDIUM CONFIDENCE

Documents containing prohibited content should be deleted or quarantined away from public access, and subsequently deleted after an agreed period (suggested six months). Content matching these sub-categories is proposed for batch sampling prior to deletion or quarantine, as there is a medium level of subjectivity in the confidence identifying it as ROT, with medium level of risk anticipated in impact on the business. Business user review, including sampling of results is suggested for these sub-categories of ROT. Connexus proposes that files in this category should be sampled in batches and deleted, as the risk of this category containing any files with business value is medium. If files are deleted in error, IT backups are available for restore, but the risk of this being required is low.

- REDUNDANT EMAIL ARCHIVE
- REDUNDANT FONT
- REDUNDANT HELP FILE
- REDUNDANT INTERNET FILE
- REDUNDANT JAVA FILE
- REDUNDANT PROGRAMMING FILE
- REDUNDANT SYSTEM FILE
- REDUNDANT ZERO BYTES
- OBSOLETE NAMED ARCHIVE
- OBSOLETE OLD DRAFT
- TRIVIAL GAME FILE
- TRIVIAL P2P FOLDER
- TRIVIAL SUSPECT FILM
- TRIVIAL SUSPECT MUSIC
- TRIVIAL ZERO BUSINESS VALUE

ROT - LOW CONFIDENCE

Documents containing prohibited content may be quarantined away from public access, and subsequently deleted after an agreed period (suggested six months). Content matching these sub-categories is proposed for quarantine, as there is likely to be a higher level of subjectivity in the confidence identifying it as ROT, with higher level of risk anticipated in impact on the business. These sub-categories will require lengthier business user review and may therefore be considered impractical during early stages of a cleansing project.

- REDUNDANT CAMERA PHOTO
- OBSOLETE STATIC UNUSED
- TRIVIAL CLOUD STORE

SURPLUS DUPLICATE FILES

Duplicates can be reported in both bulk and at a detailed level. Reports enable bulk duplicates to be viewed based upon file duplication or file contents duplication. A common approach to file level deduplication is to keep a master file in the most accessible location, replacing exact duplicates with an optional shortcut which points the user to retrieve the file from the master location. For example, if a file is held in a departmental share but is

also held in a more restricted area such as a user's own personal share, the file in the departmental share will become the master. Master files can also be determined automatically using a number of parameters.

SENSITIVE DATA REMEDIATION

Your organisation is likely to be incurring storage costs for information which also puts you at financial and reputational risk. Many types of financially, commercially, private and personal sensitive data records can put your organisation at significant risk and have the potential to incur unnecessary cost if left unmanaged. For example, the Information Commissioners Office has issued £7 Million in fines up to £400,000 each for breaches of guidelines relating to the protection of sensitive data over the last five years. Copyright infringement and breaches of the Trade Mark Act can also lead to ten years incarceration and an unlimited fine, which your office holders may be vulnerable to, even if they have not saved the offending content.

As data volumes grow (typically at 40-60% per year), this risk is often overlooked due to the perceived effort required to retrospectively apply good records management. Identify and locate files which contain confidential or sensitive data, review whether they need to be retained and if so, ensure they are suitably managed and protected, rather than left in open fileshares. We can help locate up to 35 general types of sensitive data, plus specific sensitive data stored as part of your own business processes and offer a clear plan of action for remediation and improved management of the relevant documents.

The following default categories of high confidence ROT rules are fully defined and tested:

- ATTORNEY CLIENT PRIVILEGED
- BANK ACCOUNT
- BIRTH CERTIFICATE
- COPYRIGHTED
- CREDIT CARD
- DATE OF BIRTH
- LOGIN DETAILS

- PASSPORT
- PROTECTIVE MARKING
- TRADE SECRET TERMS
- UK DRIVERS LICENCE
- UK NATIONAL INSURANCE
- UK NHS
- UK UTILITY BILL
- VULGAR TERMS

Although it may initially sound contradictory to the above advice to protect sensitive data, despite collectively working for one organisation with shared goals , we frequently see a separate departments (and even individual users) working independently, with local siloes of documents which are only clearly visible to a very limited number of staff. As noted, there can be very good reasons for this, but a first principle for business success is to share all information between all users unless there's a good business case to restrict it. The most basic elements to encourage sharing and collaboration are to implement consistent file naming conventions and a simple logical accessible filing structure, so it is clear to other users where documents of any type should be stored, and to clearly identify them in that location. A common mistake made by many organisations is to rush to implement a technology solution such as an Electronic Document and Records Management System, but successful implementation of EDRMS requires a clear and practical Information Governance Strategy as a solid foundation and framework for the management and use of information within the organisation. There are indeed clear benefits in moving content to a managed environment, where sensitive data can be properly managed, but if done without considered preparation, the EDRMS merely becomes a new and more expensive repository for the storage of data.

METADATA TAGGING

Users can often struggle to locate relevant files within a poorly structured fileshare, due to poor file naming conventions and a bloated folder structure. Making these files accessible again is most easily achieved by adding relevant metadata to make subsequent searches more effective, but asking users to add metadata to millions of legacy files is not a practical solution. Even if conservatively allowing just 30 seconds for a user to tag a single file with metadata, manually tagging 1 million files would take nearly five years to complete. A junior staff member employed to carry out this work could potentially tag around 20,000 files per month. Software can automate this process by applying metadata and classification rules based on a number of attributes, including file content. Results will be entirely consistent and defensible, as they will conform to predetermined characteristics, but can also be sampled or reviewed for accuracy if required.

CLASSIFICATION DESIGN

A classification is a hierarchy of information which can be used for content restructuring, the application of a retention policy or business taxonomy, determination of most appropriate handling of information assets in mergers and divestitures or for enhanced navigation of content. Determing where files should be located in a classification is achieved using various combined metadata values to prepare a virtual model of restructured data. The complex inter-relationship to define a suitable classification and the underlying metadata to support it are tasks which our consultants are renowned for completing quickly and efficiently. Testing and feedback cycles typically take the classification through a number of iterations, which can then be implemented as a new filing structure, or used as metadata for navigation.

ECM MIGRATION

An often overlooked aspect of implementing a new ECM system such as SharePoint or Office 365 is how that wonderful new piece of technology will be populated with helpful information from day one.

One approach is to lock down all archive content in legacy collections in fileshares, SharePoint, OpenText,

Documentum, etc. and forcing users to upload content manually as it is needed, but this means starting with an empty ECM and results in very slow and often poor user adoption.

At the other extreme, some organisations simply migrate everything they can, but this just transfers the current state of chaos to a more expensive repository. If no metadata is added, many of the benefits of the new ECM are lost. Switching users without offering significant benefits alienates them and also results in poor adoption.

We believe the answer is to migrate a lower volume of high value files - removing surplus content with no business value, then semi-automating the preparation of what is left by generating enhanced metadata. We offer a wide range of default categories and rules for this, but specialise in developing custom rules to map to business taxonomies, retention schedules and custom metadata frameworks.

A logical progression for many organisations is to implement an Enterprise Content Management System (ECM), which will introduce enhanced functionality for the capture, storage, management and preservation of relevant data to improve access and increase levels of staff collaboration on your data in future. The precise design and efficient population of a new ECM such as SharePoint should be carefully planned to ensure users gain maximum benefit from the new system with minimum effort. The design will reflect consistent organisational and metadata requirements, to ensure relevant content is available quickly to those who need it, whilst simultaneously securing sensitive information. One of the key barriers to user adoption is the transition from current storage, typically in fileshares, which can largely be overcome by populating the new system with a bulk migration of this legacy data. When completed manually, bulk legacy migrations are slow and inconsistent. Software based migrations are becoming far more common, as they deliver lower volumes of higher quality relevant migrated data with consistent metadata, reducing the volume-based cost and staff effort required for a successful migration.

There are many types of files which will be in an unsuitable format to be migrated to your EDRMS. We offer default categories of unacceptable data which will require remediation prior to migration to your EDRMS as follows:

- DATABASE FILE
- EXTENSION INCORRECT
- EXTENSION IS NULL
- LARGE DISK IMAGE
- OVERSIZED FILES
- PASSWORD PROTECTED.
- PATH LENGTH OVER 256 CHARACTERS

Certain files may also be marked by users in a way which denotes they require specific types of protection from cleansing or migration activities. Terminology within each organisation will be particularly variable for these classes of files, but our suggested categories are:

- DO NOT DELETE
- FOR MIGRATION
- LEGAL HOLD
- CUSTOM RECORD TYPES

Ongoing monitoring of content is essential to maintain compliance with your information governance policies. This monitoring process can be automated using software, and we offer guidance on how to implement this yourselves, or an ongoing managed service to deliver this successfully for you.

ACTIVE NAVIGATION SOFTWARE

Connexus Information Governance uses the revolutionary Discovery Center software from market leader Active Navigation to implement its services and methodology. Active Navigation's file analysis platform has been built specifically to rapidly discover and understand information compliance and quality problems in chaotic unstructured information stores. Discovery Center analyses electronic files in place to create an efficient index containing key metadata for use in a wide range of information governance scenarios from content clean-up and disposal, through compliance and migration to continual content governance. Connexus Information Governance adds proven methodologies and expertise creating bespoke rules packs and which are used to create specific solutions for each customer.

ACTIVE NAVIGATION PRODUCT OVERVIEW

Active Navigation Modules and Components

To allow tailoring to specific solutions, Discovery Center is modularised; modules are delivered in a single executable installation file and enabled by license configuration:

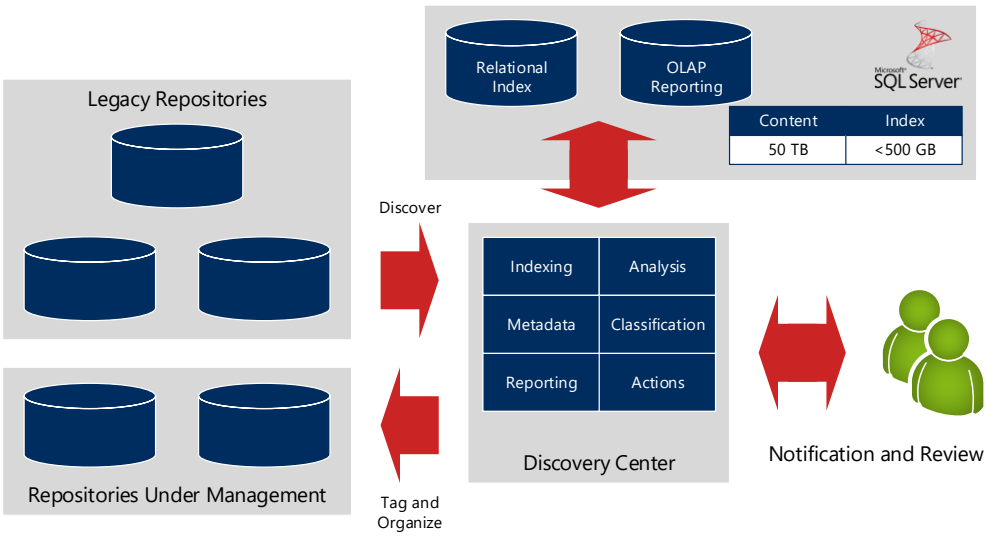
Discovery Center | Analysis | Delete and Quarantine | Tag and Organise | Connectors

Discovery Center provides capabilities from system management through metadata management to indexing and reporting. Analysis extends the indexing capabilities of Discovery Center, adding analytics of file contents whilst the actions enable the user to Delete and Quarantine or Tag and Organise files in bulk. Finally, connectors enable Discovery Center to work across a range of different information repositories.

Discovery Center is supported by the Discovery Center Workbench client application for the design and modelling of classifications. Workbench is usually installed for a small number of specialist users.

Active Navigation Technologies

Active Navigation's Discovery Center is built upon Microsoft technologies, exploiting the capabilities of Windows Server, Internet Information Services (IIS) and SQL Server to deliver a high performance and scalable solution that can be readily deployed and integrated into any enterprise environment. Windows Server 2008R2/2012/2014 provides the IIS web application host operating system and drives all discovery and analysis capabilities through the Discovery Center web application. SQL Server 2008R2/2012/2014 supports the Discovery Center OLTP database for the scalable storage of system configuration settings and index results. SQL Server Analysis Services deploys an OLAP cube for high performance reporting which provides web-based charting for visualization of discovery and analysis results. Discovery Center is developed using .Net Framework 4.5 with C# and JavaScript.



Simple Active Navigation Architecture

Active Navigation System Performance

Active Navigation is designed to index, analyse and act upon large numbers and volumes of files across any organisation's network infrastructure. Like all similar technologies, performance is largely governed by the performance of the underlying network and the specification of host hardware in use.

Scenarios for well configured systems and quality networks can achieve file discovery rates of between 500,000 and 1,000,000 files per hour (or up to 15TB per 24 hr period).

ACTIVE NAVIGATION DISCOVERY CENTER

Discovery Center provides a platform for all file analysis and reporting capabilities. It is delivered as a browser-based web-application through Microsoft's IIS. Using Windows Integrated Security it controls and provides access to all other applications and modules based upon four fundamental user roles ranging from the system administrator to a reviewer. The main functions of the Discovery Center are indexing (including network discovery), metadata management, reporting and system administration.

Active Navigation Indexing

Index Processes

The process of locating containers and files and recording their properties is known as indexing. The Discovery Center controls and manages all indexing activities and provides an administration interface for scheduling and control. Its integrated connector framework is designed to index any electronic information store using these three modes of operation:

- **Discovery (Skim)**. Launched from a specific location on the network map (such as a server or share), a skim collects file properties, such as location, name, ownership and size from the files contained within. Since the skim works only with file properties, it proceeds rapidly, handling hundreds of thousands to millions of files per hour.

- **Duplicate Analysis**. Extending the skim, a 256-bit SHA-2 hash value[1] is calculated for each analyzed file to identify duplicates. Duplicate files are clustered together in a duplication report based upon that hash value. Duplicate analysis can also identify Microsoft Office files that differ only by their extended metadata so that comparisons can be made between SharePoint and file share environments.
- **Textual Analysis**. Enabled by the Analysis Pack, analysis retrieves skimmed files from the target location and collects a wide variety of metadata extracted from their contents. Since analysis requires the extraction of all file contents, the process runs more slowly than a skim, dependent upon exact index configuration and network performance.

The following table summarises the types of properties and analyses supported by each type of index processes:

Index Process	Retrieved Properties and Analysis Types
(Discovery) Skim	File properties, repository and related container hierarchy
Duplicate Analysis	File hash and file content hash for duplicate analysis
Textual Analysis	Themes, summaries, similarity, keywords, extracted text patterns

Index Storage and Management

Discovery Center employs a Microsoft SQL OLTP database to store the results of network discovery, skim and analysis. This database is optimised for the indexing process and, importantly, does not record the full text of files. This approach stores file properties and analysis results and ensures that the footprint of each Discovery Center and its supporting database remains as small as possible, occupying disk space between 0.1% and 5% (typically less than 1%) of the volume of indexed information.

The Discovery Center provides full control over all indexing options; these include as the selection of analysis types, the allocation of index credentials, creation of index to metadata mappings, incremental indexes and index scheduling.

Information Sources

Discovery Center's Connector Framework provides a series of connector modules to enable the indexing of different information sources. Connectors enable consistent handling of different sources while the Connector Framework enables file properties and metadata to be mapped into a homogenous metadata schema.

Active Navigation Metadata Management

Discovery Center supports a fully customizable metadata schema which draws together skim and analysis results with a powerful classification engine in order to consistently label (or tag) any indexed files. Typical uses for such metadata include:

- Labeling files based upon their value or risk to the organisation.
- Applying record or content types for migration to SharePoint or other content management systems.
- Adding equipment or case number metadata for search and process automation.
- Re-organizing files for migration to a new file plan or similar file structure.

[1] SHA or Secure Hash Algorithm is a cryptographic hash function designed by the National Security Agency (NSA) and published by the National Institute of Standards and Technologies (NIST) and is a Federal Information Processing Standard (FIPS). A hash function is a method of encoding a message. SHA-2 is considered "collision free" and is ideal for encoding documents.

Active Navigation Classification

File properties, folder structures and analysis results provide a rich range of facets for all indexed files. Classification brings these facets together into user-defined hierarchies using customizable rules with reflect information and/or business policies. The results are stored as user definable metadata. Whilst it is possible to use file classification for almost any conceivable application, common uses include:

Classification Uses
Identifying redundant, obsolete and trivial (ROT) files for cleansing.
Profiling files for triage that contain sensitive information according to the risk they present to the organisation.
Labeling files with content types to support migration to SharePoint.
Classifying and tagging files against a business classification scheme or taxonomy.
Determining file location for migration in a new file plan.

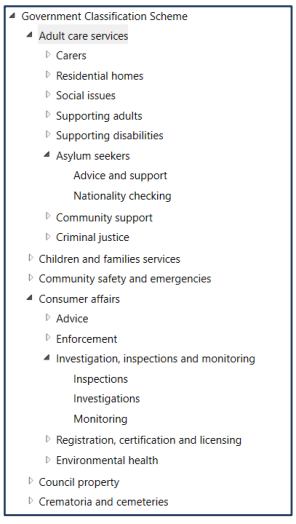
Classification rules use Boolean operators and expressions, such as ranges and similarity, to group files according to a wide range of customizable facets:

Group/Class	Attribute/Metadata
Dates and age	accessed_date, created_date, modified_date, age_by_accessed, age_by_created, age_by_accessed
File properties	extension, file_format, file_size
Location	path_length, folder, path
Access and ownership	owner
Status	retrieval_status, thematic_status
Thematics*	sentence, theme
Markup value	has specific value, has any value
Office file, repository and extracted metadata	has specific value, has any value, count of unique hits, count of all hits
Extracted pattern value	has specific extracted value, has any extracted value

*Thematic analysis is described later in this document under ACTIVE NAVIGATION TEXTUAL ANALYSIS.

Discovery Center Workbench

The Discovery Center Workbench provides a powerful visual tool for the user to explore the facets extracted from analyzed files in order to design hierarchical classification structures for metadata or the design of folder structures for green field environments. Existing structures or taxonomies can also be imported for refinement where appropriate. The design process creates a set of nodes within a structure, each supported by a Boolean rule used to match files to the node according to their metadata field attributes.



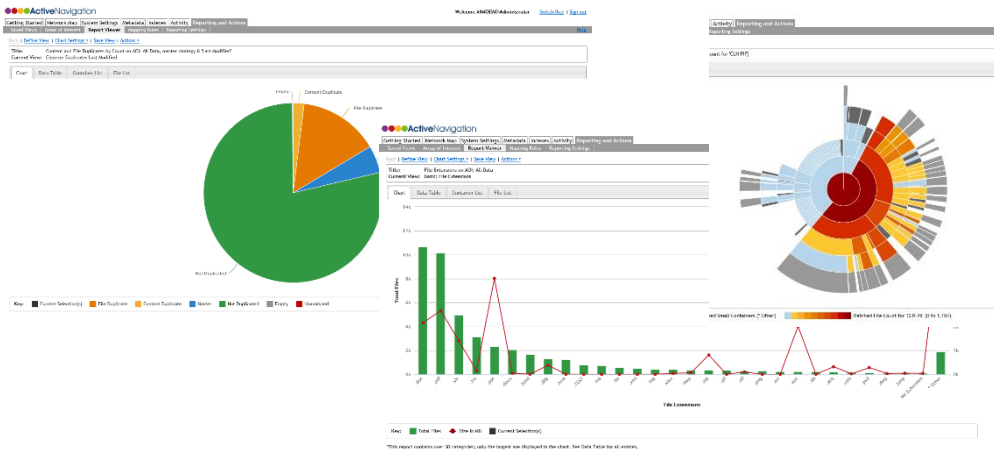
The Discovery Center Workbench has several key features to simplify structure design and support its dynamic refinement. Facets collected from files is displayed along with their prevalence so that frequently-occurring values may be extracted to build

rules for nodes. Coverage analysis classifies files in any selected information set against the structure being designed so that the success of the design can be validated. Files not classified are displayed as orphans so that their metadata values may be used to enhance node rules.

Active Navigation Reporting

High Volume Report Views

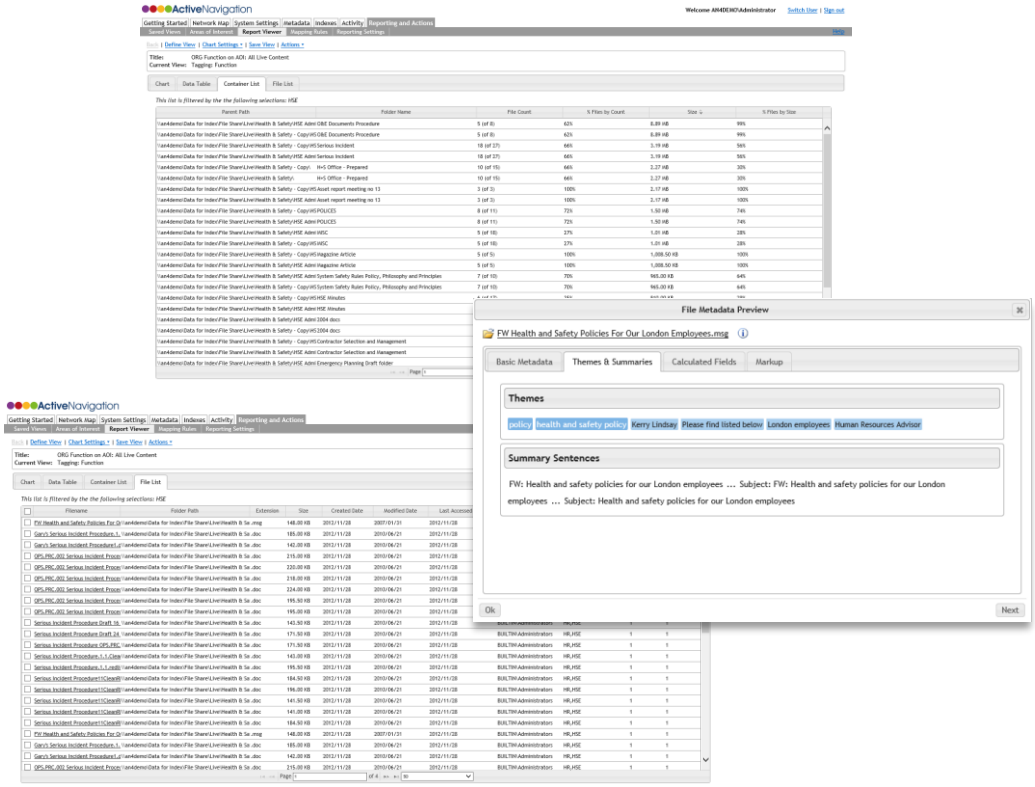
Discovery Center uses Microsoft SQL Server Analysis Services to provide report views across 100s of millions of files to allow information managers and reviewers to explore any content or metadata attribute across any connected repository. All reports can be filtered by any metadata attribute or file property and report views can be saved for re-use as required. Default reports are provided as follows:



Default Report Type	Default Report Features
File Extensions	All discovered extensions including files with no extension.
File Ownership	NTFS 'file owner' and 'created by' properties for connected repositories
File Dates	File date properties 'date created', 'date modified' and 'date last accessed'.
File Ages	File age properties calculated from each of the above file dates.
Container Usage	Folder/site/library sizes and file counts including summary information on largest/smallest and empty folders.
Duplicate Files	Duplicate files and duplicate file contents based upon SHA256 hash results.
File Types	Based upon either the file extension or file identification algorithms, files are grouped into customizable types such as 'Spreadsheet', 'Word Processing, 'Raster Image' etc.
File Size	Based upon file size properties, files are grouped into customizable size bands.
Incorrect File Extension	Files with incorrect extensions by cross-matching extensions with file identification results.

Container and File Level Reporting

To complement file overview reporting any chart or table can be drilled into in order to view individual files and the containers (folders, sites or libraries) within which those files are stored. This powerful facility provides critical information about the context of those files to inform decision making and accelerate review and action. Further, it allows the exploration of file properties in detail to both verify report contents and explore more detailed analysis results such as themes, summaries or extracted metadata.



Areas of Interest

By default, reports allow the exploration of indexes based on a single location such as a folder, SharePoint site collection or a folder. However, Discovery Center also supports the creation of reports using virtual locations known as 'Areas of Interest' to combine the results of any indexed location into a single report. This allows for reports to be configured to reflect, for example, business units or teams that have their information spread across several different servers or even mixed between SharePoint and a file share.

Active Navigation System Administration

System management features include user security, role mapping, definition of system constraints, permissions and credentials management, maintenance of the network map and connector configuration. All are controlled via the Discovery Center browser interface.

ACTIVE NAVIGATION TEXTUAL ANALYSIS

Discovery Center, provides a wide range of capabilities to extract value from the text contents of files. The results of these analyses are used to derive metadata and build a rich picture of information value, risk and quality for indexed files.

In order to analyze files from a range of proprietary formats, the Discovery Center uses Oracle's industry standard OutsideIn conversion libraries supporting over 450 unique formats to convert file contents to HTML prior to analysis. The full range of supported formats is listed on the Oracle OutsideIn homepage.

Discovery Center's analyzers fall into two categories:

- Text analyzers employ thematics and pattern matches to interpret file text contents.
- File analyzers interpret key file elements to determine specific features such as identity and file hash value.

The results of each analysis process are then made available as metadata for tagging files on migration, for reporting to explore file features or for classification to enable information profiling or re-organisation.

Active Navigation Text Analysers

Theme Extraction

Thematic extraction algorithms use a proprietary hybrid linguistic-statistical process to identify the most relevant concepts within a file's contents. A combination of language knowledge and text structure is used to identify candidate themes within the file.

Files are first processed to identify their primary language and this information determines the appropriate approach for interpreting content, including stemming algorithms and rules for decomposing natural language elements. File structure is then parsed in order to identify content in terms of words, punctuation, sentences, paragraphs and other constructs. This information is utilized by the thematic analysis processes, in conjunction with the language-specific rules, to identify theme and summary information.

Statistical comparison of the candidate themes is used to estimate the relevance of extracted themes relative to each other; this relevance ranking allows themes to be sorted and the best themes are then returned as a result of the analysis. Options for controlling the choice of themes returned include maximum number of themes, percentage of candidate themes, filtering by score, and filtering of sub phrases.

Summary Generation

Base text analysis identifies sentences as one of the fundamental constructs within file contents. In parallel with the thematic extraction process, the significant content of sentences is collated so that sentences can be compared and ranked. A summary of the file is generated by collecting the highest ranking sentences in the order in which they occurred.

Language Support

Thematic analysis is driven by a set of language packs which support document language detection based upon statistical occurrence of common short words. The detection process then selects the corresponding language pack with which to perform analysis. Language support is available for English, French, German, Italian and Spanish.

Keyword Extraction

Where an organisation employs a fixed set of keywords (such as in an existing business classification scheme, taxonomy or thesaurus), Discovery Center can be configured to extract those terms as part of its text analysis. Keyword definitions include support for word phrases and synonyms as well as case control and flexible matching of word separators.

Text Pattern Extraction

Text patterns occur independent of language and can frequently signify important information features such as project codes, social security numbers or case identifiers. Regular expressions may be defined for this purpose and can be freely customized to match any pattern of interest. Matched patterns are stored as part of the indexing process for later use as metadata or for file classification.

Active Navigation File Analysis

File Identification

Oracle's OutsideIn technology is used to determine the true format of more than 450 files formats regardless of other features such as file extension or name. File identification can differentiate between files saved in different versions of the same application (for example, Microsoft Word 2007 and Word 2010). The file identity is stored for later use.

File Properties

Property analysis allows file properties to be extracted from structured formats (e.g. Office formats 2003 to 2010) or from extended file system properties (e.g. as held in NTFS Alternate Data Streams) including EXIF properties for images.

Active Navigation File Scoring

When analysis has been completed, Active Navigation generates intensity and diversity scores for files based upon analysis and classification results. Scores are made available to reporting so that very important, very risky or very low value files can be quickly identified, sifted out from the rest and acted upon.

Active Navigation Conditional Analysis

When analyzing large volumes of content, conditional analysis allows filters to be applied to more appropriately focus network and hardware resources. These conditions select which files are brought across the network to Discovery Center for analysis and can significantly increase effective analysis performance, especially in challenging network conditions.

ACTIVE NAVIGATION ACTIONS

All reports available through Discovery Center provide facilities for action. Filters can be applied according to pre-determined user definable rules or in a free-form manner in order to focus in on specific information features. In this way, low value files can be cleansed, files containing sensitive information can be isolated and high-value files can be collected together for migration. All actions are implemented using the APIs provided by the connected source.

Active Navigation Available Actions

Discovery Center supports a comprehensive set of file actions for the transformation of files for a wide range of use cases. Available actions are:

- **Delete/Quarantine**. Deletion collects selected files together and removes them from the indexed location, optionally capturing copies in a configured quarantine location. Deleted files not quarantined cannot be recovered. When duplicated files are deleted, shortcuts can optionally be created to the chosen master file.
- **Move/Migrate**. The migrate action collects selected files together and moves them to a new location creating new container (folder) structures according to a wide range of customizable rules. This enables files to be re-organized based upon:
    - A flat (no folders) structure to leverage metadata-based navigation.
    - A rationalized/collapsed version of the original structure with options to replicate permissions within file shares.
    - A new structure based upon a defined metadata field. Using Active Navigation's classification features, such a field could replicate, for example, a new file plan or classification scheme.
- **Markup**. When reviewing files, users use markup fields to apply user-defined labels or tags as part of a structured process to support, for example, a large-

scale cleansing or migration activity. Special manual markup fields are provided so that values remain persistent regardless of indexing and analysis processes.

- **Update Metadata**. For files stored in repositories such as SharePoint, the Update Metadata action writes new metadata to files in place so that fresh or updated metadata values can be consistently applied without the need for user interaction.
- **Export**. The attributes of all selected files can be exported into a CSV or XLSX file. File exports include file UNC paths and user-selectable index attributes and analysis results to support offline review or for upload to third-party applications.

These actions are supported by a range of options to help optimize content including replacing illegal characters in file and folder names, writing shortcuts to help users locate migrated files, removing containers that are left empty after the action and writing new metadata to the target destination.

Common uses for actions include:

Action Uses
Cleansing or archival of ROT™ files.
Isolation of files from different business functions according to business or information management policies.
Re-organisation of files into new file plans.
Identification and collection of case files spread across different storage locations.
Isolation of files containing risky or sensitive information.
Migrating files to a connected repository (such as SharePoint or a records management system) with new metadata.

The details of all actions are recorded in a comprehensive audit log, available for download, showing all exceptions and successes along with details of the rules and options used for the action and the user that committed the action.

Active Navigation Metadata Mapping and Migration

To support migration to SharePoint or other connected repository (such as a records management system), the Discovery Center provides comprehensive metadata mapping facilities to control how analysis results are matched to destination metadata fields. Migration leverages metadata derived by indexing analysis and classification in order to populate metadata fields with newly derived values. Features support:

- Population of destination metadata fields, including SharePoint content types and managed metadata. Mapped metadata derives its values from Active Navigation analysis and classification rules to support a wide range of business applications.
- The removal and replacement of illegal characters according to customizable character mapping rules.
- Application of content types based upon metadata field values.
- Where appropriate, the creation of document libraries and child folders using standard move/migrate options.

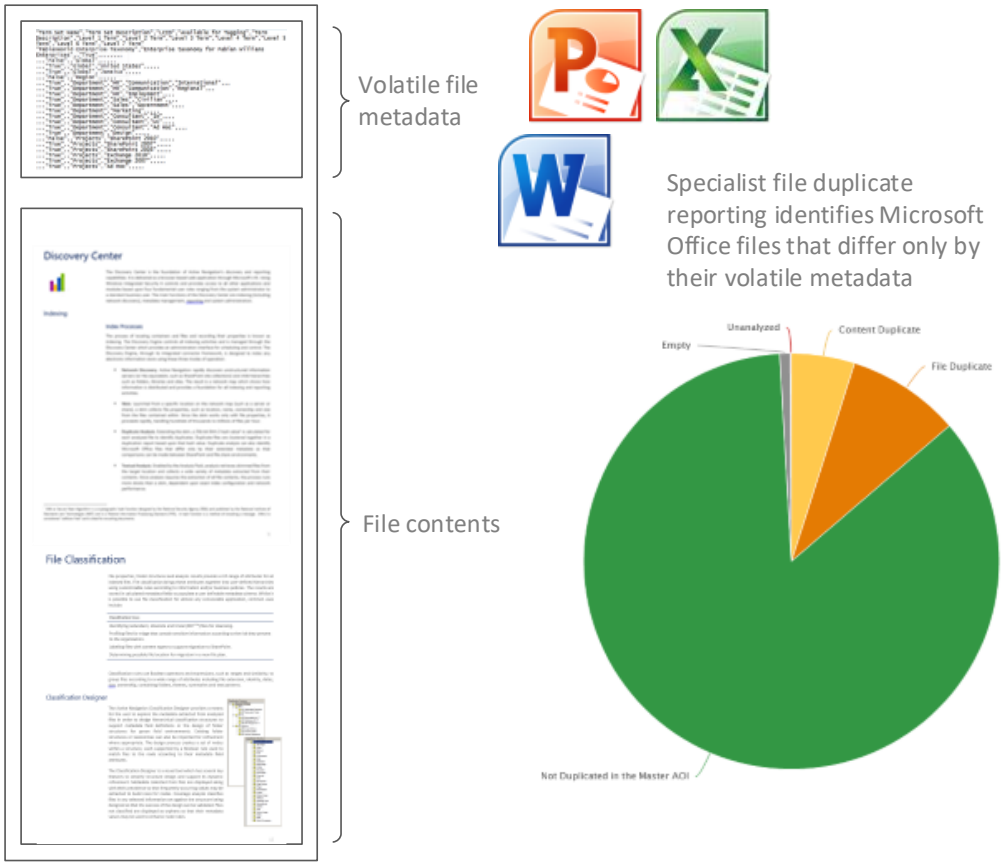
Where files already exist in SharePoint and new metadata needs to be applied, the Update Metadata function makes this possible so new relevant metadata can be applied to indexed content.

Active Navigation Duplicate Cleansing

Duplicate Analysis

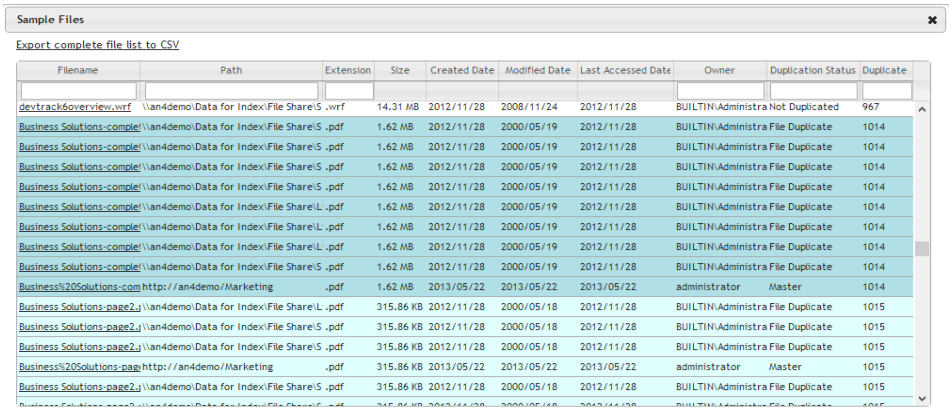
Discovery Center provides specialized features for the handling of duplicates both in bulk and at a detailed level. Reports enable bulk duplicates to be viewed based upon file duplication or file contents duplication as follows:

- File duplicates compare binary data for entire files using a SHA256 algorithm. File duplicates are exactly the same on a bit-by-bit basis.
- File content duplicates specifically ignore volatile elements of compound Microsoft Office 2003 to 2013 formats. File content duplicates have the same contents but their metadata values differ – this occurs when a file, for example, is copied to SharePoint. File content duplicate analysis allows files in file shares and SharePoint sites to be compared and duplicates identified.



Duplicate Reporting and Cleansing

The Discovery Center reporting interface provides facilities to allow the effective cleansing of both file and content duplicates. Cleansing is achieved by defining a rule which clearly identifies master files in any cluster of duplicates. Master files can be defined according to their modified date, created date or location. Using a location-based master rule (defined by an Area of Interest) enables authoritative record sets, or 'Golden Corpuses', to be defined and records leaked from them removed to ensure a single source of truth is maintained. Master and surplus files are clearly identified and can be reviewed before cleansing takes place. During cleansing, shortcuts can be optionally created in place of surplus files so that users can readily location the remaining master.



Active Navigation Automation and Review

Report Views and Policy Definition

As part of a repeatable defensible process for content remediation and disposal, Discovery Center enables re-usable views to be defined to capture and report upon content responding to a specific set of metadata rules. Where those rules reflect content policies, views allow policy violations in any connected repository to be quickly identified and remediated after indexing has been completed.

Notifications

Whenever a report view has been defined and saved it is possible to allocate users view so that they will be automatically notified if any view contains responsive content. Notifications can be controlled within user profiles to appear within the Discovery Center interface and/or to send an e-mail to the assigned user(s).

In this way, views reflecting content policies can be continually applied to any connected repository as part of a scheduled indexing process and responsible users notified should file be found that respond to those policies.

Work Packages

Where closer control is required of review processes, Discovery Center supports the definition of work packages. Work packages assign a bundle of reviews to users from report views. When a work package is sent it describes the review work required, the deadlines for the review and the actions required as a result. As work packages progress, the work package owner is notified of progress and may review a log of actions to date.

Activity History and Audit Trail

Whether as part of a structured work package or more ad-hoc use, Discovery Center maintains an activity history which captures all content indexing, review and actions. Each entry in the activity history captures relevant context including report types and selections made as well as actions taken. Further, the results of any action are supported with a file-by-file audit trail log of all successful actions and exceptions.

ACTIVE NAVIGATION CONNECTORS AND REPOSITORIES

Discovery Center works with a range of repositories to support content discovery, cleansing, migration and the ongoing application of policies for content governance. Its Connector Framework provides a set of interfaces which enable, through the provision of connectors, any potential information source to be analyzed and the resulting basic file properties and calculated metadata mapped into a single metadata schema. The results can then be leveraged to report and action based upon a consistent view of the health and condition of information in any connected source. This consistent view provides the foundations for information governance and a powerful source of metadata for cleansing and migration into a single target destination.

Repositories are supported as follows:

- **Close Integration**. A specific connector exists which addresses unique needs and features of the repository. Such connectors may include components that are installed on the repository host environment to access advanced features.
- **General Integration**. A connector exists that addresses the repository in a generalized way based upon an industry standard protocol such as the Common Internet Filing System (CIFS) or CMIS (Content Management Interoperability Services). Functionality may have minor limitations but offer adequate capability without the need for close integration.
- **Process Integration**. A process is available which uses the functionality of both Discovery Center and the target repository to achieve some integration between the two. Process integration supports a wide range of repositories and provides an efficient and cost-effective approach for specific use cases, most often for the purpose of content migration.

Existing Support	**Repository/Environment**
Close Integration	Microsoft SharePoint 2003, 2007, 2010, 2013 Microsoft Exchange 2010, 2013 Office 365 SharePoint and Exchange
General Integration	Network file shares including Windows, Novell, NetApp and any CIFS-compliant target Cloud gateways such as Commvault, StorSimple and Riverbed Whitewater
Process Integration	Alfresco One, EMC Documentum, HP Record Manager, OpenText Content Server, RSD Glass
Planned Support	**Repository/Environment**
General Integration	CMIS-compliant repositories including Alfresco One, EMC Documentum, HP Record Manager, OpenText Content Server

ACTIVE NAVIGATION MANAGEMENT REPORTING

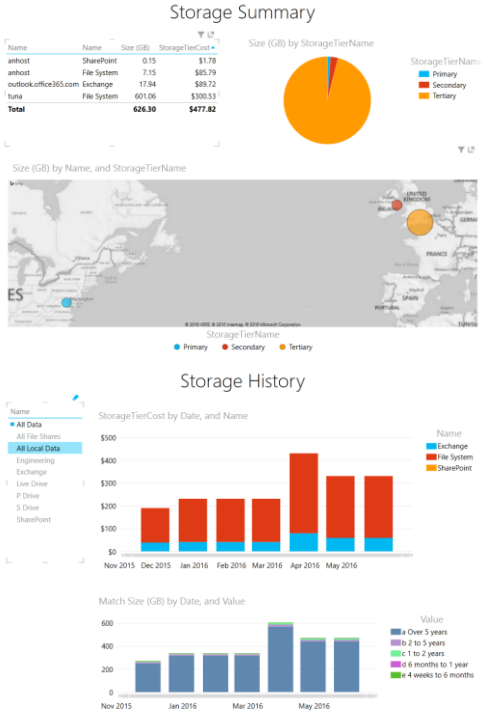
Whilst Discovery Center provides comprehensive reporting for review and action down to file and metadata detail, its Management Reporting module creates and maintains aggregated data specifically for the creation of overview dashboards and reports. Management reporting data includes:

- All volumes and file counts by index and area of interest.
- Storage cost metrics by configured storage tier.

- Break down by discovery center instance, connector and repository.
- Mapping of physical repository server to geographic location.
- File volumes and counts by selected calculated field values (or policies).
- File volumes and counts for indexing, action and migration.

The above metrics are collated on a daily basis or as requested and filterable as snapshots for daily, weekly, monthly, quarterly or annual reporting, proving a historic reporting record.

Finally, for deployments across geographically-dispersed organisations or large volumes of content, customers who choose to deploy additional discovery centers can connect to one management reporting database for aggregation across all content and locations.



Management reporting data is designed to be used with customers' existing business intelligence and reporting tools and is ideally suited to the many business applications, including:

- Tracking storage return on investment for information projects.
- Creating project dashboards for stakeholder buy-in.
- Matching information policy successes and violations to specific functions, departments or stakeholders.
- Tracking progress and success for key information policies over time.

CONNEXUS
INFORMATION GOVERNANCE

08458 675 516
www.connexus.consulting

ABOUT CONNEXUS INFORMATION GOVERNANCE

Connexus Information Governance offers expertise in the field of information management strategy, delivery and training, with collectively over 50000 hours experience. As a preferred service provider of Active Navigation in Europe, we offer unparalleled expertise in delivering implementation, training and consultancy on projects using Active Navigation software. Founded by Chris Dewey, the company provides a network of exceptional consultants who know how to address the issues and pitfalls in information governance, having led, managed, planned and delivered projects of all sizes across the globe.