SLIM Applications

SharePoint - Extract metadata, find duplicates, and delete at scale

BulkMetadataManager is a command-line application for SharePoint administrators. It runs on a Windows system in your own environment and works with SharePoint Online, SharePoint 2019 and SharePoint Server Subscription Edition.

One tool, three modes:

- **Extract metadata** – extract properties from emails, Office files, PDFs, images and videos into SharePoint columns, without a new version and without changing Modified / Modified by.
  Simultaneously process different content types with different file types.
- **Detect duplicates** – inventory of documents, store details in a local SQLite database, and find candidates (content hash, Message-ID, and related signals). Phases can run together or separately. Support for incremental runs.
- **Bulk delete** – read document URLs from a csv file and delete in per-site batch requests. Default: recycle bin (typically restorable for 93 days). Optional: permanent delete.

BulkMetadataManager empowers organizations to fully leverage the capabilities of the SharePoint platform, enhancing user productivity and adoption. Utilizing metadata in SharePoint columns allows for efficient filtering, sorting, grouping, and powerful searches. Additionally, it helps organizations meet compliance requirements and keeps storage consumption under control.

Key Benefits:

- Enrich documents in SharePoint with metadata
- Search, filter and retain original 'document' dates – not the upload date
- Make email attachments searchable
- Find true duplicates (including emails) instead of relying on file name alone
- Delete at scale with logging; recycle by default so mistakes are recoverable
- No external SaaS: the application and database stay on a machine you manage

How It Works:

- **Metadata** – download (CAML) → extract → update columns.
- **Duplicates** – inventory → SQLite → compare → csv file/report.
- **Delete** – csv file with URLs → group by site → delete to the recycle bin or permanently.

BulkMetadataManager features

Functionality

- Extract metadata from common file formats such as msg, eml, docx, xlsx, pptx, vsdx, pdf, jpg, png, dng, heic, heif, tiff, jfif, cr2, mp4, mov and wmv.
- Simultaneously process different content types with different file types
- Detect duplicates across the entire tenant using a sophisticated algorithm.
- Bulk delete documents.
- Make email attachments searchable by indexing their contents.
- Support for viewing email threads in SharePoint by extracting Message-ID, References, In-Reply-To and Normalized Subject.

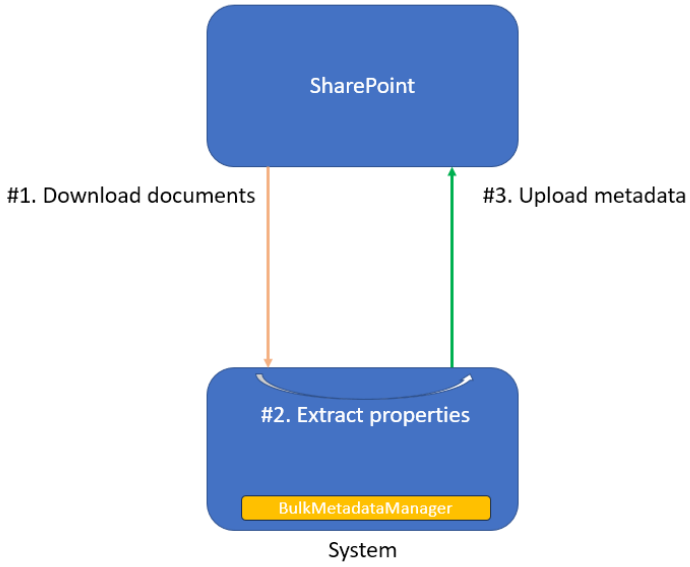
Technical Capabilities

- High performance via parallel processing and optimization techniques.
- Supports horizontal and vertical scaling.
- Automatic handling of SharePoint Online throttling.
- Works with large lists exceeding the 5,000-item threshold.

Deployment

- Install the executable on a separate machine.
- Configure metadata mapping.
- Run via command line or schedule as a task.

SLIM Applications



Metadata Extraction Capabilities
BulkMetadataManager supports comprehensive metadata extraction across a wide range of file types:

- **Emails**: Extracts up to 30 properties, including *Subject*, *From*, *Sent Date*, *Attachment Names*, Message-ID, Conversation GUID and more.
- **Office Documents**: Supports extraction of 29 properties, such as *Title*, *Subject*, *Company*, *Last Modified*, etc.
- **Binary Office Documents**: Support for extraction of 23 properties.
- **PDF Files**: Retrieves 19 metadata fields, including *Title*, *Keywords*, *Modified Date*, and others.
- **Images and Videos**: Extracts properties like *Date Taken*, *GPS Coordinates*, and other embedded metadata. Support for common metadata schemes like EXIF, XMP, IPTC and ICC.

These capabilities allow organizations to enrich their SharePoint libraries with meaningful, searchable metadata—automatically and at scale.

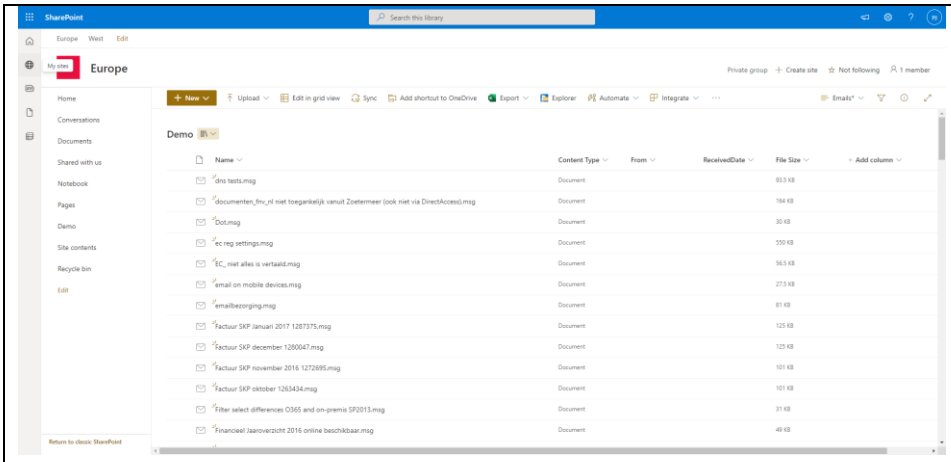
Before and After: Email Metadata Example
The following figures illustrate the metadata of an email document before and after processing with BulkMetadataManager:

Supported Environments

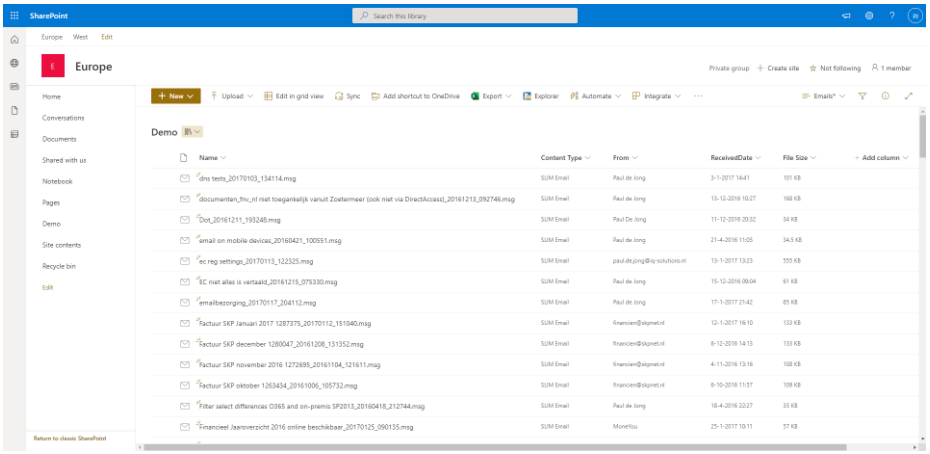
- Compatible with SharePoint Online, SharePoint Server SE and SharePoint 2019.
- Works with documents uploaded via web, OneDrive, mobile, or external apps.

Demo
Interested in seeing it in action? Contact us to request a live demo of BulkMetadataManager.

Licensing
Licensed per tenant.



After:



Please note that SharePoint views may only display a subset of the metadata added by BulkMetadataManager. In addition to enriching metadata, the application also renames email files and updates their content type to ensure consistency and better classification.

Make Email Attachments Searchable

By default, email attachments are not always indexed by SharePoint Search, creating blind spots that hinder both user productivity and regulatory compliance.

BulkMetadataManager addresses this limitation by extracting the contents of common attachment formats—including pdf, docx, xlsx, pptx, vsdx, xml, csv, txt, and zip - and making them searchable within SharePoint.

This unique capability enhances the discoverability of critical information and supports compliance with privacy, legal, and data governance regulations

Duplicate detection

Using BulkMetadataManager with the 'Mode=Duplicates' builds an inventory of documents (path, size, SHA256 hashes and email identifiers such as Message-ID) in a

SQLite database created on the same system as the BulkMetadataManager application. A second phase compares candidates using a sophisticated algorithm:

- For documents unaffected by property demotion (pdf, jpg, mp4, …) the SHA256 hash is used. Special logic is applied to large files to minimize download the full files.
- BulkMetadataManager automatically handles documents affected by property demotion like docx, pptx, …
- For emails the unique value of the MessageID is used.

Use 'Mode=Duplicates' to run the full process or run inventory and comparison separately. Set 'LastDuplicatesRunDate' to process only documents changed since the last run.

By default, the PreservationHoldLibrary is excluded. This library can be included by using setting the property 'IncludePreservationHoldLibrary=true'.

The output is a list in csv format for further analysis. The csv file can be used for the Delete mode.

Bulk deletion
'Mode=Delete' reads URLs from a csv file. Requests are grouped by site and sent as SharePoint batch operations.

Default: deleted items go to the recycle bin. With standard first- and second-stage settings they remain restorable for 93 days.
Permanent delete: 'UseRecycleBin=false'. Log lines are marked "Permanently deleted".

Retention labels, records and "eDiscovery holds" still block deletion; those items are logged as failed, not removed.

Delete requires write permission on the target sites (for example 'SharePoint: Sites.ReadWrite.All', or 'SharePoint: Sites.Selected' with the Write role). Read-only app permissions are enough for inventory and detection.

Logging and Traceability
BulkMetadataManager generates multiple log files to support transparency, troubleshooting, and auditing:

- **Summary Log**: Provides a high-level overview of the processing results, including the number of documents processed, success/failure counts, and key actions taken.
- **Detailed Log**: Captures granular information for each individual document, including extracted metadata, processing status, and any encountered issues.

- **Duplicate logs**: this log is created for Mode=Duplicates and Mode=DuplicatesInventory.

These logs help administrators monitor operations, verify outcomes, and ensure accountability in large-scale metadata management tasks.

Performance and Scalability
BulkMetadataManager is engineered for high performance and efficiency, even in large-scale environments:

- **Parallel Processing**: Utilizes multiple parallel threads to download documents and apply metadata updates, significantly reducing processing time.
- **Smart File Handling**: For extraction of metadata from large video files, only a small portion (e.g., 4 MB of an 800 MB file) is downloaded to extract metadata, minimizing bandwidth usage.
- **Caching and Optimization**: Implements intelligent caching and other performance enhancements to streamline operations.

Scalability and Resilience
Horizontal and Vertical Scaling: supports both scaling out (adding more instances) and scaling up (increasing resources) to boost throughput while avoiding SharePoint throttling.

- **Throttling Management**: Automatically detects and handles SharePoint Online throttling. The application pauses and resumes operations based on the Retry-After period, ensuring uninterrupted processing.
- **Large List Support**: Fully compatible with SharePoint lists that exceed the 5,000-item view threshold, enabling reliable operation in enterprise-scale environments.

Data Integrity and Preservation
BulkMetadataManager ensures that your document history remains intact during metadata updates:

- **No Version Changes**: Metadata is applied without creating a new document version.
- **Preserves Audit Fields**: The original **"**Modified" and "Modified By" fields in SharePoint are left unchanged, maintaining the authenticity of document timelines.

This approach guarantees that metadata enrichment does not interfere with existing audit trails or version control policies.

WhatIf Mode: Safe Testing Without Impact
The WhatIf option in BulkMetadataManager allows you to test your configuration and validate metadata extraction — without making any changes to SharePoint.

- **No Metadata Written**: SharePoint columns remain untouched during the test run.
- **CSV Output for Review**: All extracted document properties are logged in a separate **CSV file**, enabling detailed analysis and verification before going live.

This feature is ideal for trial runs, audits, and configuration tuning, ensuring confidence and control before applying changes in production.

Deployment and Automation
BulkMetadataManager can be executed via the command line or configured as a scheduled task, enabling seamless integration into automated workflows.

This flexibility allows metadata to be added to documents regardless of how they were uploaded—whether through the SharePoint web interface, OneDrive for Business, mobile clients, or external applications.

- The configuration is controlled via the BulkMetadataManager.ini file.
- A CAML query provides fine-grained control over which sites, libraries, or documents are processed.

Upcoming Enhancements
Future releases of BulkMetadataManager will include:

- **Importing metadata from CSV files**
- **Document labelling** for retention and sensitivity policies

These enhancements aim to further expand the tool's capabilities and compliance support.

Target Audience
BulkMetadataManager is designed for use by:

- **SharePoint Administrators**
- **Site Collection Administrators**

It is not intended for end-users, ensuring that only authorized personnel manage metadata operations at scale.

SLIM Applications
"Light-weight applications with rich functionality"
info@slimapplications.com
www.slimapplications.com