DATASHEET

migration center

The machine learning improved migration process

How migration-center's Amazon Comprehend integration can help to increase the metadata quality during your next content migration

Complex migration projects expose different challenges – one of them refers to the metadata or classification of unstructured content like Microsoft Office, PDF documents, or images. Those metadata are often incomplete, incorrect, or simply do not exist in the source systems. During the migration process, much more effort is needed to fix incomplete metadata because it must be added manually, or external data sources are required to complete the missing information.

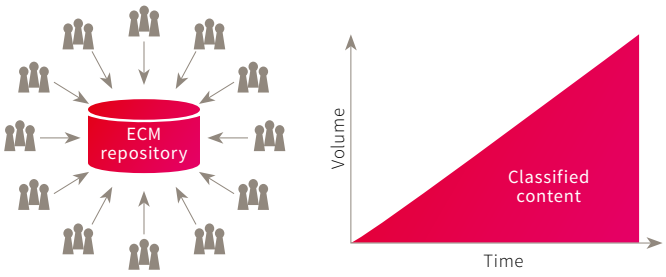
aws partner network

Select Consulting Partner

Why integrating Amazon Comprehend matters to create meaningful metadata

However, enhancing the unstructured content with meaningful content is highly recommended to sustain the integrity of the application's data. At this point, machine learning based natural language processing comes into play – a machine learning supported classification process can reduce this effort dramatically.

Growth of classified content within an ECM system



Many companies run ECM applications for decades and over these entire years, users typically create terabytes of manually classified documents in existing ECM systems. This classified content is a real treasure chest for a migration project and natural language processing is the right technology to dig up this treasure.

The Amazon Comprehend integration of migration-center allows users to leverage the power of natural language processing and machine learning in order to discover insights from content files and generate meaningful metadata for their next content migration.

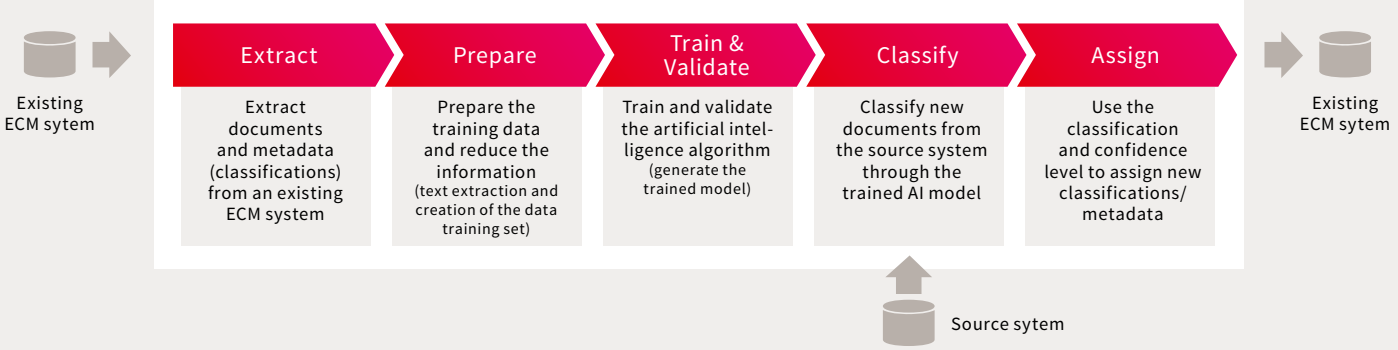
Key capabilities of migration-center's Amazon Comprehend integration

- **Out-of-the-box integration with Amazon Comprehend**
- **Endpoint support for Entity Recognition, Language Detection, and custom classifiers**
- **Multi-metadata recognition**
- **Analysis of documents (Word, Excel, PowerPoint, PDF etc.) with OCR**
- **Text extraction from image formats and/or embedded images within documents**
- **Full integration in the migration process**
- **Full traceability of the migration process for a legal and audit-compliant migration**
- **Simulation mode to support Amazon Comprehend service cost calculation/estimation**

Amazon Comprehend Services

Based on 6+ years of AWS consulting experience, fme is able to build architectures for various use cases including enterprise level architectures. In addition to our professional migration services offerings, we would be happy to help you to build Comprehend or Comprehend Medical models for your domain-specific needs and support you on your way to becoming a confident Comprehend user.

Phases of the migration-center Amazon Comprehend integration for individual classifiers



Fundamentals

The Amazon Comprehend service contains a lot of built-in classifiers like Entity Recognition and Language Detection, named standard classifiers. Users can work with the following existent classifiers out-of-the-box if they meet the requirements:

- **Entity Recognition**
  The Entity Recognition returns the named entities (»People«, »Places«, »Locations« etc.) that are automatically categorized based on the provided content
- **Language Detection**
  The Language Detection automatically identifies content written in more than 100 languages
- **Amazon Comprehend Medical**
  Amazon Comprehend Medical is a natural language processing service to extract health data from doctors' notes, clinical trial reports, and patient health records

Amazon Comprehend also offers high flexibility by allowing users to create their own custom classifiers and custom entity recognizers. With a set of specific migration-center features, users are now able to extract the existing data (metadata plus content) from an ECM application to train an individual Amazon Comprehend classifier (see figure "Phases of the migration-center Amazon Comprehend integration").

The resulting trained classifier can then be used to automatically classify documents with incomplete or missing metadata during the migration.

The individually trained custom classifier will categorize the metadata specific for user's domain or industry based on the text content of a document.

Afterwards, the user can define the threshold of acceptance based on a confidence value between 0 to 100 %. If the chosen confidence level of an attribute has been reached, migration-center automatically sets this value as a new attribute for the corresponding document.

Possible Use Cases

Merge ECM repositories and enrich metadata
When merging several ECM repositories, documents of the same category might have different attributes. Instead of assigning the missing values manually, the existing values can be used to train a classifier that assigns the most suitable value to each missing attribute automatically.

Improve eDiscovery by extracting information from the unstructured content
Finding information often requires extensive search and well-consistent metadata. migration-center's Amazon Comprehend integration supports the process of creating metadata from unstructured content and filing new documents into an eDiscovery platform.

Reduce the effort of creating meaningful metadata
In a survey of 1.500 office workers, 93 % answered to be unable to find documents because they are badly tagged and 82 % said it would be beneficial if the system automatically tagged documents.[1] The Amazon Comprehend integration assigns meaningful metadata to new documents in an ECM repository.

Increase the quality of search results
Users lose their patience quickly when filtering relevant search results by themselves. Therefore, documents with many correctly tagged attributes improve their search experience significantly and allow them to focus on more crucial tasks.

Archive file shares to an enterprise archive together with a context
Moving outdated files from a file share to an enterprise archive often requires additional metadata, which don't always exist. The Amazon Comprehend integration automatically predicts missing metadata.

Onboard file shares to an ECM system and set meaningful metadata
Importing data from a file share might include a laborious process of labeling attributes manually. With a trained classifier, attribute values can be assigned to documents easily while migration-center migrates documents from the file share to the ECM repository.

fme

fme group | Germany · Romania · USA
T +49 531 238540 · info@fme.de
T +1 203 6174250 · info@fme-us.com

[1] *The 2019 Intelligent Information Management Benchmark Report, Publisher: M-Files*