DATASHEET

migration center

The artificial intelligence improved migration process

How artificial intelligence and migration-center can add additional value during your next content migration

Complex migration projects can be very painful and expose many challenges. One of such challenges refers to the metadata or classification of unstructured content like Office or PDF documents. Those metadata are often incomplete, incorrect or simply do not exist in the source system. To fix incomplete metadata during the migration process, much more effort is needed because metadata must be added manually or external data sources are required to complete the missing information. To protect the investment made in the ECM application and to sustain the integrity of the application's data, it is highly recommended to go the extra mile and enhance the unstructured content with meaningful metadata. At this point artificial intelligence (AI) comes into play; an AI support classification process can reduce this effort dramatically.

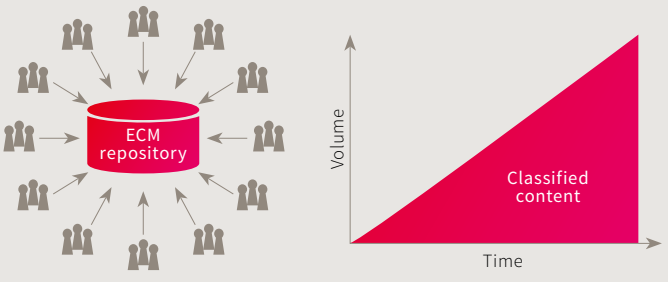
Many companies run ECM applications for decades and over these entire years, users typically create terabytes of »manually« classified documents in existing ECM systems. This classified content is a real treasure chest for a migration project and AI is the right technology to dig up this treasure.

The Auto Classification Module is a highly-parameterizable software that can be used as a plugin for migration-center. The module can be used from within migration-center and third-party software.

Fundamentals

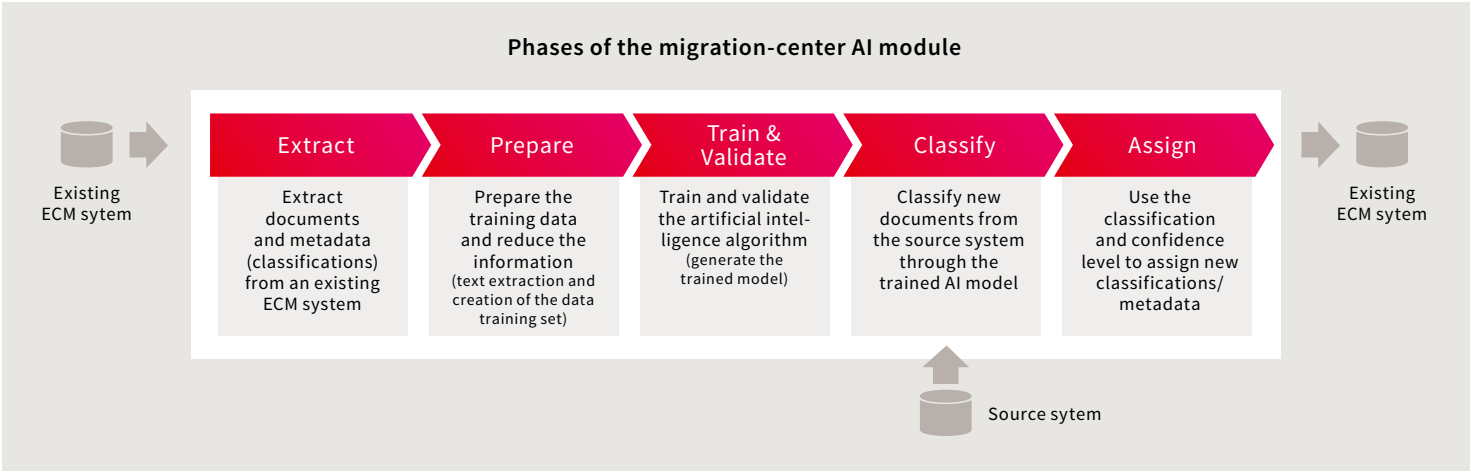
At fme we introduced capabilities to our product migration-center, which allow us to use the existing classified documents as training data for our new AI based Auto Classification Module. With this set of features, users are able to extract the existing data (metadata plus content) from an ECM application and train an AI algorithm (see figure »Phases of the migration-center AI module«). The resulting trained AI model can then be used to classify documents with incomplete or missing metadata during the migration automatically. Our algorithm detects patterns based on the text content of a document and refers their patterns to possible attribute values. Afterwards, the user can define the threshold of acceptance based on a statistical confidence value between 0-100 %. 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.

Growth of classified content within an ECM system



Additional benefits

- **Merge ECM repositories and harmonize/enrich metadata:** When merging two or more ECM repositories to a single repository, documents of the same category can have different attributes. Instead of assigning missing attribute values manually, the existing attribute values can be used to train a classifier. Then while merging the repositories, the trained classifier can automatically assign the most suitable value to each missing attribute.
- **Improve eDiscovery by extracting information from the unstructured content:** Finding important information for any legal case requires extensive search and well-consistent metadata. An AI classifier can support the process of creating metadata for unstructured documents and filing new documents into an eDiscovery network.
- **Enhance security: Set confidentiality classifications for documents:** In every company there are different levels of confidentiality. Freely available, personnel matters or strategic documents of the board of directors are just a part of conceivable levels. Often it is not simple to distinguish between levels. An AI model can help to propose the confidentiality of a document.
- **Reduce the effort of creating meaningful metadata:** In a survey of 1.500 office workers, 93 % say that they are unable to find documents because they are badly tagged



when filed and 82 % say that it would be beneficial if the system they use automatically tags documents.[1] With the help of migration-center, the Auto Classification Module can assign consistent metadata to new documents in an ECM repository.

- **Analyze the quality of existing metadata within an ECM system:** In the previous use case, it is mentioned that 93 % of office workers cannot find documents due to badly assigned metadata. An AI classifier can validate existing metadata. If the existing metadata does not match with patterns of similar documents, it can propose new values.
- **Increase the quality of search results:** A user's search experience is heavily influenced by relevant results. Users lose their patience fast and it consumes a lot of their time to filter relevant documents by themselves. Therefore, documents with many correctly tagged attributes can improve a user's search experience tremendously and allows them to focus on their more crucial tasks.
- **Archive file shares to an Enterprise Archive together with a context:** Moving outdated files from a file share to a dedicated enterprise archive often requires supplying additional metadata. In many cases, metadata does not exist or employees intuitively assume it, while they browse the file share. The Auto Classification Module can automatically predict missing metadata.
- **Onboard file shares to an ECM system and set meaningful metadata:** Similar to merging multiple ECM repositories, importing data from a file share often includes a labor-intensive process of labeling attributes manually. Training a unique AI classifier with attributes from the ECM system makes it possible to assign attribute values to unstructured documents easily. Assigning these values happens on the fly while migration-center migrates the documents from the file share to the ECM repository.

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

Technical features

The Auto Classification Module is equipped with several parameters and filters to be able to create a unique classifier for a specific use case. Every classifier exists of a data transformation pipeline and algorithm. The following list represents a short summary of possible transformation and filter options:

- **Multi-output classification:** Often it is desirable to predict more than one attribute. The Auto Classification Module supports the prediction of multiple attributes with only training a single classifier.
- **Optical Character Recognition (OCR):** OCR can be performed on scanned documents or embedded images within documents. A neural network analyses the images and converts the text content to uniformed text. To adapt to accuracy, specific hardware and performance requirements, the DPI and color scheme of the analyzation can be defined.
- **Parsing of Microsoft Office files, PDFs and images:** A vast variety of file types can be consumed. With the help of OCR, even scanned documents can be processed.
- **Text reduction and clean up:** Many documents contain a large text corpus. By determining which words have a special meaning for a single document and within a dataset, it is possible to reduce text to the most meaningful words. Furthermore, many words in a natural language, like English or German, do not add any informational gain and appear very frequently. These so called stop words can be removed from a text corpus automatically.
- **Algorithm support:** Scientific literature developed multiple algorithms in the last years. Most of them have a specific purpose. Therefore, the Auto Classification Module supports three different algorithms to provide the best performance and accuracy for different use cases, hardware and data sizes.

fme

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