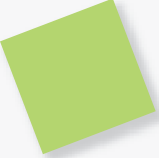




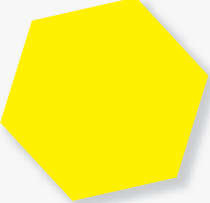
Easy content migration with EntropySoft Content ETL Migration Methodology



Increasing volumes of unstructured data need to be moved across disparate systems. Typical migration cases include:

- transferring data from a legacy system to a new system
- publishing information across systems
- archiving information to records management systems

In this document, EntropySoft presents a simple migration methodology that maximizes the use of Content ETL and therefore simplifies transfers and migration. Content ETL has two clients; Content ETL Studio which is used to design graphically the different processes and manage migration jobs and Content ETL Web that could be used for web execution.



This document is based on EntropySoft's experience of various large scale migrations done using the Content ETL product. These migration guidelines can be applied to any of the 30+ content-centric systems for which EntropySoft provides read / write connectors.

This methodology may not be exhaustive and only reflects EntropySoft's point of view on automated high-volume migrations. We are not concerned here with detailed planning, budgeting and internal allocation of resources required by the migration.

EntropySoft's migration methodology can be separated in five successive phases.

ACTIV™ Methodology

- 1- Analyze and build
- 2- Content-transfer design
- 3- Test
- 4- Implement
- 5- Verify




1. Analysis and build

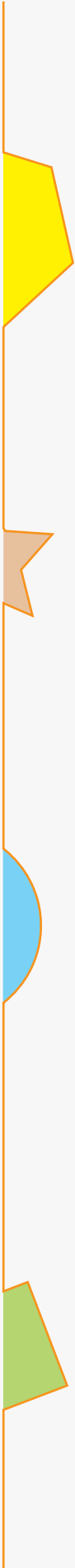
1.1 Analysis of source content

This phase is crucial to enumerate exactly all objects in the source repository and what needs to be migrated.

The source analysis must include all information on document types and attached metadata, as well as the total volume of information to be transferred.

A non exhaustive list of relevant information includes:

- 
- The document types
 - The number of documents according to type
 - Attached metadata according to document type
 - The presence of annotations
 - The presence of multiple versions
 - Permission heritage from parent folders
 - Errors and locked documents
 - Non-typical documents (because of size for example)



Content ETL wizard can generate automatically a process that builds a descriptive report of the environment that has to be migrated. This analysis process can be adapted easily by drag & drop to cover custom requirements or to take in consideration the source inherent properties.

For very large migration, even the enumeration of these properties can be laborious. This analysis process can be modified to reflect different crawling strategies for faster results.

The analysis process identifies each particular document present in the source in order to build a representative sample. This representative sample should be used in the next phase to validate with the customer the migration process.

Information pertaining to environmental conditions that impacts the overall process must also be collected at this point:

- resources capacities (migration speed)
- systems availability (migration schedule)
- possible document use and modifications during the migration (content transformation)

Phase 2 should take in consideration these requirements. For instance, if users continue working on the source application, multiple iterations must be anticipated.

1.2. Implementation of the target repository

At this point, the target application usually needs to be manually installed and configured. This is not the aim of this methodology to explain this phase. This is usually a separate project.

2. Designing the content-transfer process

The designing and the testing phases (phase 2 & phase 3) can be done several times until the process is ready to be used and validated.

The main objective of this designing phase is to build one or more repeatable processes that will achieve the content migration. A process is a series of n stages that handle specific action on the document. The elementary object that crosses over the process is called a parcel. A parcel includes the document plus all related information (versions, permissions, annotations...). The design of these content processes must include all stages that answer the customer requirements.

The Content ETL “new project” wizard helps designing a working process that just need to be parameterized. This wizard guides the user through different options that take in consideration different requirements (users and groups mapping, permissions mapping, Item types mapping and a few other...).

Once the raw process is built, custom requirements are covered by a palette of pre-programmed stages or the ability to implement custom stages. This building flexibility and openness is one of the main strengths of Content ETL.

Content ETL Studio includes a high number of pre-programmed stages that can be used to complete the process. These stages are added with a simple drag & drop.

These pre-programmed stages can be completed by custom-designed stages. Content ETL offers an effective way of creating new custom stages. This mechanism is really simple to use and allows unlimited content or metadata manipulations.

The different content-processes can of course be saved so as to be duplicated or re-used as templates and possibly enriched with new stages for the forthcoming migrations.

All pre-programmed stages can be easily accessed on the palette on the right hand side of the screen.

Help can be accessed through the f1 key.

Examples of pre-programmed steps include:

- Pdf rendering
- Branching and conditions
- Automatic mapping of source and target folders
- Automatic mapping of permissions
- Automatic mapping of users and group
- Automatic mapping of items type.

Custom-designed steps can include:

- Timestamp
- Metadata completion with customers' own structured data.
- Metadata cleansing
- Document deduplication.

3. Testing the content process

Once the target repository is deployed and a first draft of the process is ready, the testing and debugging phase can start.

A test migration is to be started on the representative sample of all types of documents, including problematic documents, heritage issues, oversized files etc... This representative sample is built by Content ETL during the discovery phase. Content ETL Studio is used to start the first transfer.

The user can set breakpoints on each of the stages of the process. Once the job is started in debug mode, as soon as the first parcel reaches a breakpoint, Content ETL Studio opens a debug mode perspective where the user can drill down into the parcel.

This perspective gives an access to the different metadata and all the information available in the parcel. Then, the process can be run step by step. This allows in-detail checking of the validity of the content design and process.

Iterative testing is done to debug and optimize the content process.

The roll back of migrated object can be easily achieved within Content ETL studio.

During this phase, the architecture and multithreading adjustments need to be considered for optimizing performances.

The dashboard indicators of Content ETL Studio help automatically determine the resource-consuming steps. Thanks to the individual stage multithreading option, a higher number of threads can be allocated to the most consuming tasks.

The dashboard presents all the metrics related to a stage. It is therefore possible to monitor the number of parcel per second, ...

As soon as the representative set of document is correctly migrated with the best performance, the project team including customer representatives must check the target application to certify that all requirements are verified.

Based on the results of the pre-migration phase, an agenda for the final migration can be defined. This can be put to action thanks to EntropySoft Content ETL job scheduling option.

For large migrations that need to be distributed to power users in the organization, Content ETL Studio allows the deployment of the process on the Content Federation Server. We will explain in the next phase the main benefits of using this web execution.

4. Implementing the migration

The implementation of the migration is usually highly depending on the customer requirements. Sometime, the customer asks for a big-bang where during a short period of time the transfer has to be executed; sometime, users continue working on the source applications and this implies the scheduling of multiple iterations. One other aspect to consider is the migration type. Depending on the source and target applications, some migrations can be achieved in one run. Then, some migrations need to be divided into several smaller migrations.

Elementary migrations jobs can be scheduled easily on a regular basis including job stop and job start. These jobs can be manually or programmatically started.

Each job can be run multiple times until the completion. Content ETL stores the previous runs status and migration of failed objects can be retried quickly. Each one of these successive migrations is validated by intermediary reports, which can list replica ID, failures and many more information.

Content ETL offers strong web execution capabilities. Any users in the organization can help moving the documents he is in charge of. Even if Content ETL Studio is commonly used to manage migration, in some cases, it is crucial to be able to spread the role to functional users. Therefore Content ETL offers a web access that allows users to browse the source and the target applications and to manage the transfers. All the monitoring capabilities of Content ETL Studio are available in Content ETL Web. A report is sent via email when the job terminates. Then errors can easily be managed using the Content ETL web access. Jobs can be restarted to manage remaining errors.

5. Verifying the migration results

After each migration, a comparison test between source and target is automatically generated by EntropySoft Content ETL.

EntropySoft recommends that a separate software program does the final check on the exhaustiveness a of the migration.

