

The Definitive Guide to Automating Content Migration

Migrating digital content without scripting or consultants.

September 2012

Contents

- 1 Executive Summary
 - 2 Content Migration: The Hidden Challenges
 - 3 Traditional Content Migration
 - 4 Content Migration with Kapow Katalyst
 - 7 Kapow Katalyst Content Migration in Action
 - 8 Automate Your Content Migration with Kapow Katalyst
-

Executive Summary

Today's organizations are adopting Content Management Systems to reduce cost, increase efficiency, and improve customer service, but migrating digital assets into a new CMS can be a significant effort.

Too often organizations begin a migration without fully understanding the challenges they face—resulting in slipped deadlines, reduced scope, and even project failure.

This white paper will alert you to the key issues that can arise when planning and executing a CM project, and show how Kapow Katalyst™ has addressed those issues—in project after project—and enabled cost and schedule reductions up to 90%.

Kapow Katalyst enabled Lenovo to access content held in a diverse group of data stores, including Quest DB, Domino DB, static HTML, and custom .NET applications.



Content Migration: The Hidden Challenges

Today's organizations depend on sharing a wide variety of digital content, including websites, digital media, and documents. But selecting the right Content Management System to house and manage your digital assets is only the first step—the real challenges arise while migrating content to your new CMS. What makes content migration (CM) so complex, unpredictable, and high-risk? Let's explore a few reasons.

Digital Assets May be Missing

It's not uncommon for assets to be lost or misplaced. Some assets may be held in file systems, or located in the cloud on sites such as Flickr, YouTube, or other hosting sites—beyond control of the source CMS. Missing assets—which show up as broken links—will require additional, unplanned effort to locate, capture, and migrate.

Content Restructuring is Difficult

Nearly every CM project includes restructuring of digital assets caused by merger, acquisition, business or product reorganization—and often this seems straightforward. However, technical scope often expands as business requirements are better understood, causing changes to ripple through an entire project—and increasing project cost and risk.

Dynamic Pages and Assets Are Challenging to Migrate

Migrating static content is challenging enough, but migrating dynamic pages or content raises complexity to a new level. Web 2.0 techniques such as JavaScript that dynamically alter web pages during rendering defeat migration techniques that rely on static page images. Pages with dynamically-created assets, built on-the-fly from database applications emitting HTML or XML, also need special methods.

Metadata Needs to be Created and Migrated

Having complete and accurate metadata is especially critical for a successful migration. For example, Web page metadata carries the keywords that guide search engines: without the correct keywords content will be harder to locate and the web site's search position can suffer. But this metadata is scattered in multiple places. Some is in the HTML code used to build a Web page, some may be created dynamically, and some may need to be extracted from the asset or even inferred based on asset content or attributes. Wherever it may be located, capturing, analyzing, and transforming metadata generally adds significant effort.

Resource Links need to be Reassigned

Websites work because of the links that exist between digital assets, but migrating links during migration is not simple. Your migration process will first have to capture all of the old links from the existing site to preserve its structure. Then, because each asset has a new and different link assigned to it by the new CMS, your process will have to revisit old pages and content and update them to use the new links. This two-stage process is complex during even a simple migration, and is especially challenging whenever content is restructured—and if not performed correctly, the new site will not function as planned.

Traditional Content Migration

Considering the above obstacles, it's not surprising that CM projects have gained a reputation for being late, over budget, and disruptive, both to the organization and to its customers. Until now there have only been two options for CM projects: migrate assets manually, or develop custom-coded migration scripts.

Option 1: Manual Migration

The first approach is to “cut, paste, and hope,” using teams to manually copy data, and save it—one asset at a time—directly into the target CMS or a staging area. This approach is wholly inadequate for all but trivial amounts of content. Besides being far too slow, the errors it will generate cannot be discovered until after the target CMS has been populated, days or weeks later—and each error will need to be researched and corrected manually, dramatically increasing cost and extending the schedule.

Option 2: Services/Scripting

The second approach is to develop a custom set of scripts in an attempt to:

- Extract all of the assets, including metadata and links;
- Transform it into the new CMS format and restructure it as needed;
- Load it into the new CMS and reassign all links.

These custom software development projects demand specialized skills that are likely to require high-priced staff augmentation from consulting firms—and no real assurance that the project will be completed on a reasonable schedule.

Content Freeze: a “Best Practice?”

It's commonplace for migration projects to prohibit updates to the source content—this makes it unnecessary to detect and synchronize changes to migrated assets. The “content freeze” has even been elevated into a “best practice” even though it's far from ideal. Freezing content disrupts the business and introduces business risk, and traditional migration techniques provide no support for eliminating content freezes during your project.

Trial Migrations: Unsupported

Performing a trial migration is essential, but manual migration provides no support for this best practice and scripting doesn't do much better. Both methods exhibit the same significant weakness: the quality of the migration cannot be verified until long after an error has been introduced. Without strong support for automation, many projects simply “patch” errors in the target environment, using ad hoc, non-repeatable efforts, causing timelines to grow unpredictably.

Too often, large migrations become “big-bang” projects, with lengthy content freeze periods and many error correction cycles.

In the worst cases the project fails.

The Association of American Medical Colleges needed to create metadata from the thousands of PDFs they were migrating to their new CMS, and migrate existing metadata as well. They succeeded with Kapow Katalyst



The U.S. Navy migrated content from thousands of websites onto a single CMS platform, using Kapow Katalyst to restructure content.



Content Migration with Kapow Katalyst

Content Migration with Kapow Katalyst is fundamentally different from other approaches. Rather than relying on manual effort or custom development, Kapow Katalyst provides a proven process built on a complete environment designed to support your CM project. Its built-in automation addresses the challenges implicit in CM and eliminates the risk and expense of using obsolete approaches.

Step 1: Automated Inventory Analysis



Figure 1. The proven, automated, Kapow Software Content Migration Process.

Unlike other solutions, Kapow Katalyst addresses the first, critical step of any CM project: developing a complete inventory of all digital assets before the migration begins. To do this, Katalyst relies on one of its unique features: a patented extraction and analysis capability that automatically and iteratively crawls 100% of your website. Within hours you will see a detailed analysis of your website—including the type, amount, and current location of the assets to be migrated, and the consistency of the HTML and quality of the links it contains. This automated service has saved thousands of hours of effort for our customers and enabled them to resolve significant issues before the project starts.

Step 2: Automated Migration

With a complete, accurate inventory in hand you can proceed with your CM project. Our proven process enables you to migrate and test your content quickly and reliably, with a minimum of disruption to your organization and your customers.

Universal Extraction

Kapow Katalyst performs the following for each page in a website:

- Collect metadata from the web page (title, creation date, URL, author etc.)
- Collect HTML content sections and remove headers and footers
- Loop through all images and documents (PDF, PowerPoint, Excel etc.) and capture them

Katalyst can loop over multiple pages or repeating elements in a single page, and it can choose alternative paths depending on page content—even dynamically created content—thanks to its integrated JavaScript engine. Katalyst sees all dynamic activity as the page is built, and not just the fully rendered version.

Kapow Katalyst can extract all data and metadata associated with any object visible on the website, even if the content resides in off-premise hosting services. Katalyst can easily extract complex data objects such as:

- **HTML Content:** Stores HTML content, URL, and metadata in the HTML page such as keywords
- **Binary Content:** Stores documents, their URLs, the URLs of the linking pages, and any additional metadata contained within the document, such as the author of a PDF document
- **Links:** Stores every URL on each page, and is used later in the process—as content is loaded—to update content links to point to their new locations
- **Email:** The structure of emails can be captured and transformed into other document types
- **Collaborative spaces (calendars, folders, blogs):** The content and structure can be captured and mapped onto corresponding structures in the target CMS, even without native support
- **Breadcrumbs:** Kapow Katalyst can map existing structures such as breadcrumbs into folder structures

It's not necessary to model extracted data in advance because Katalyst's content data models can be refined and extended as new pages, data types, and content types are discovered during the course of the project. And the modeler is integrated—there's no need to leave Katalyst as additional types of data are encountered in the source.

Instant Error Correction

You control Kapow Katalyst extraction with an integrated IDE called the Design Studio, which accesses your source website in real time. You specify extraction rules as you build each automated migration workflow, and you see the results immediately. Any errors can be corrected in seconds, and you don't even leave the page until the extracted data is correct. This cuts days or weeks from correction cycles, and is a major factor in the 90% reduction in schedule that our customers routinely report.

Powerful Transformation

Once data has been extracted from the page, Kapow Katalyst cleanses and transforms it according to the rules you specify in Design Studio. Katalyst includes a wide variety of built-in transformations you can use without coding. With Katalyst you can easily convert HTML to semantic XML, or to RSS-compliant XML for syndication—or transform dates such as “two hours ago” or “yesterday” into usable data. Virtually any transformation rule you need is built into Kapow Katalyst—including the ability to invoke Web services, Java, and .NET applications if necessary.

To support migration projects that incorporate a staging area (often used to simplify content reorganization) Katalyst can automatically define, create, and update relational tables in any major SQL database—without writing SQL. Or, if you prefer, you can create and manage XML—or update a standard file system. You get all the benefits of instant error correction while transforming data, further enhancing productivity.

Intel tried cut-and-paste migration but it took them an hour per page. After 2 years they called Kapow Katalyst, and migrated their remaining 300,000 pages in only 8 weeks.



Kapow Software offers Content Migration solutions for leading CMS systems including:

- OpenText
- EMC Documentum
- Oracle WebCenter
- HP Autonomy
- IBM
- SharePoint
- Hippo
- Tridion
- Sitecore
- EPiServer
- ... and more

The website for the American Society for the Prevention of Cruelty to Animals includes database search, behavior advisors, and multiple content types—all migrated by Kapow Katalyst.



Flexible Loading

Once transformation has been completed, Kapow Katalyst provides three methods to load the new CMS:

- 1. Load content via the CMS' GUI:** With this method, you build workflows that log in and complete the web forms that add each asset to the CMS. This mimics the behavior of a human user—except this method is 100% repeatable, and can be performed in a fraction of the time.
- 2. Use the CMS vendor APIs:** You build workflows that drive vendor APIs using Web services (SOAP or RESTful)—or generate Java or .NET code to invoke vendor APIs. Workflows can also generate XML or any other format as input to a loader.
- 3. Store the data in a database:** If the target CMS supports direct database access, Kapow Katalyst workflows can update the CMS' database without the need to code SQL, all through Design Studio's point and click development interface.

Step 3: Verify the Migration—Without Freezing Content

Trial migrations are considered a CM best practice, but most CM solutions, offer no support for this step. As a result, corrections are performed in the target environment, and a content freeze may have to be imposed.

Kapow Katalyst includes an integrated scheduler that can be used to run your CM workflows in any sequence. This enables you to run most migrations in a few hours—from the very first extraction step to the final load of the target CMS—and provides support for frequent trial migrations. Katalyst's resilience to web page layout changes gives you the ability to re-run a migration quickly and accurately—and avoids the need to freeze content.

Kapow Katalyst Content Migration in Action

U.S. Department of Defense

A large DOD client chose Kapow Katalyst for their CM solution because they found no other technology could perform the migration in their required timelines. The client created an advanced content management system that incorporated collaboration tools such as email, messaging and calendars as well as more traditional CMS capabilities.

Challenges

- Time Limitations: current software was expiring
- Content imbalance: content types in source systems were not represented in the new system (for example, e-mails existed in the source, not in the target)
- Multi-level folder structures had to be retained
- Manual migration was too costly and time consuming to consider

Results

- Kapow Katalyst enabled the client to migrate the content in 1/10th the time of traditional methods
- Content types were converted during the migration
- E-mails were converted to text documents and added to an e-mail folder for each user
- Calendar entries were extracted and recreated in the target system's calendar
- Multi-level folder hierarchies were migrated

Lexmark anticipated a lengthy content freeze as they migrated 50 country-specific websites into a single instance. With Kapow Katalyst they didn't need to freeze any content—or write any migration scripts.



Automate Your Content Migration with Kapow Katalyst

Kapow Software's Content Management solution was created to address the challenges posed by traditional, manual CM processes.

The automated Kapow Katalyst process is unique. It begins with a thorough inventory of the existing content, enabled by Kapow Katalyst's unique ability to view all web page activity and content—even for pages built dynamically.

An integrated development environment is used to specify even the most complex extraction and transformation rules without the need to develop any code. These rules are tested and debugged in real-time, increasing accuracy and shrinking effort to a minimum.

An integrated scheduler allows the entire migration process to be rerun on demand, avoiding the need to freeze source content. Kapow Katalyst can also access any web site that requires authentication, automatically completing web forms and navigating through multiple web pages whenever necessary.

Taken together, these factors have enabled Kapow Katalyst Content Migration to demonstrate reductions as much as 90% in cost and effort compared with manual migrations or custom-coded solutions.



Kapow Software
260 Sheridan Avenue, Suite 420
Palo Alto, CA 94306
Phone: +1 800.805.0828
Fax: +1 650.330.1062
www.kapowsoftware.com
sales@kapowsoftware.com

About Kapow Software

Kapow Software is a leading innovator in the integration market with its intuitive, powerful application integration platform that harnesses the power of cloud, mobile, social and big data without the need for consultants, coding or APIs. Its Integration as a Self-Service™ platform is the first of its kind to empower employees and IT leaders to collaborate in new ways to help modernize the workplace, increase agility and improve business results. Kapow Software is trusted by more than 600 large global enterprises, including Audi, Intel, Fiserv, Deutsche Telekom and more than a dozen federal agencies for process automation, application integration, big data collection, mobile enablement, content migration, web data intelligence, social media monitoring and other mission-critical solutions. For more information, please visit: www.kapowsoftware.com.

Kapow Software and the Kapow Software brandmark are trademarks of Kapow Software. All other product or service names mentioned in this document may be trademarks of the companies with which they are associated.