



Automating Content Migration Using Software, Not People

Copyright 2008 Kapow Technologies

All rights reserved

<http://www.kapowtech.com>

 **kapow**
TECHNOLOGIES

Contents

Introduction.....	3
Content Migration Challenges.....	3
The Manual Approach	3
The Coding Approach.....	3
The Kapow Approach.....	4
Example Use Case	4
Content Modeling	4
Content Collection	5
Content Transformation	6
Content Loading.....	7
Testing, Debugging and Deployment	7
Results.....	8
Other Customer Examples.....	8
U.S. Department of Defense	8
U.S. Health and Human Services	8
Conclusions	9
Content Migration Cost Comparison.....	10
More Information.....	11

Introduction

Kapow Technologies software allows content to be migrated with unprecedented speed and accuracy. Major Federal agencies and corporations such as U.S. HHS, U.S. DoD, EMC, and Intel use our software to achieve as much as 80% time and cost savings compared to traditional approaches for content migration. Kapow software can migrate content from and to content management systems (CMS) from major CMS vendors such as, EMC/Documentum, Vignette, Stellent, Interwoven, IBM Workplace, Microsoft CMS, Tridion, SiteCore, Lotus Notes, as well as custom solutions. We have partnerships with many of these organizations. This white paper discusses Kapow's software approach to content migration as it compares to other content migration methods.

Kapow Technologies

Kapow Technologies offers a family of software products that provide business solutions for content migration, portal enablement, lightweight integration/web service enablement, or on-demand harvesting of web content. For purposes of this paper though, we will be focusing on our Content Migration solution.

The Kapow Technologies software platform comes with a visual point-and-click development environment, a stateless runtime server, and management and monitoring tools. Using the visual development environment, one can quickly create "robots." A robot is visual script that implements a set of instructions that creates interactions with any combination of web based applications, databases, web services; RSS/Atom feeds, and file formats such as XML, PDF and Excel.

In this whitepaper, we've included two customer examples that illustrate the benefits of using Kapow to implement content migration, with proven cost savings of over 70% compared to more conventional approaches.

Content Migration Challenges

Migration of content between CMS systems from different vendors, or even between different releases from the same vendor, is a daunting task. Even a project of moderate scope typically takes many man-months of effort and costs hundreds of thousands of dollars, with more complex projects easily costing in excess of a million dollars. There are two approaches to content migration generally practiced today.

The Manual Approach (cut and paste)

- Time consuming
- Prone to human error
- Costly

The Custom Coding Approach

- Major coding effort
 - Time consuming
 - Costly
- Moving custom content types is difficult
- Maintaining content-links is difficult
- Metadata within HTML content is difficult to extract and is usually ignored

Both of the above approaches are slow and have serious accuracy issues due to human error. In addition, most CMS's are so customized in the actual implementation that vendor-provided API's cannot be used as a vendor-provided solution for content migration. These challenges make existing approaches to content migration a complex, costly and time consuming process.

The Kapow Approach

Our approach to content migration is fundamentally different, resulting in efficient automation of the three key steps performed during content migration:

- Collecting content from the source CMS(s) and/or target web sites
- Transforming the content
- Loading content into the new CMS

A typical robot takes just hours to build using our visual development environment and the results are much more accurate than manual content migration. In addition, Kapow can collect and analyze non-standard types of content and files including e-mail, collaborative spaces like calendars, folders and blogs and transform both the content and structure to match the new CMS's structure and content types.

Example Use Case:

Federal Government Agency's Content Migration Needs

The following is an assessment performed for a large federal government agency with a pending content migration project. We began by getting to know their source content. We built a content collection and analysis robot that ran against their .gov domain and collected HTML pages, PDF documents and images. We executed our robot against 2,844 web pages and collected HTML, PDF and binary content. We were able to collect HTML from 99% of the pages and 100% of the PDF and image content. The content source analysis gave us an idea of the type and amount of content to be migrated as well as consistency of HTML formats across the site.

Note: This was a use case intended to gain analysis of a segment of the agencies .gov web presence and is not the entire site.

The following processes were automated within hours to achieve the above results:

- Content Modeling
- Content Collection From Legacy Site
- Content Transformation and Storage into a Temporary Database
- Content Loading into Target CMS

Step 1: Content Modeling

Using our visual development environment we created a data object that modeled the content for collection. We give you the ability to model many different object types such as databases, XML, web services, web forms, and file formats. Modeling the content is a straight forward process of defining the storage fields for where you want to put extracted content. It is unnecessary to model an entire domain before beginning the migration process. The model can be refined and extended as new data, pages and content types are discovered over the content migration life cycle. Once the model was defined, a Kapow wizard created the database tables.

Content Storage and Linking Object Types:

- **HTML Content;** Stores HTML content, URL of the page and metadata in the HTML page such as 'Keywords'.
- **Binary Content;** Stores documents, their URL's, the URL's of the pages they are on and any additional metadata contained within the document such as the author of the PDF document.
- **Links;** Stores every URL on each page. During content loading this is used to update the content linking within documents to point to their new locations.
- **Errors;** If anything goes wrong, the URL of the page and error messages will be logged.

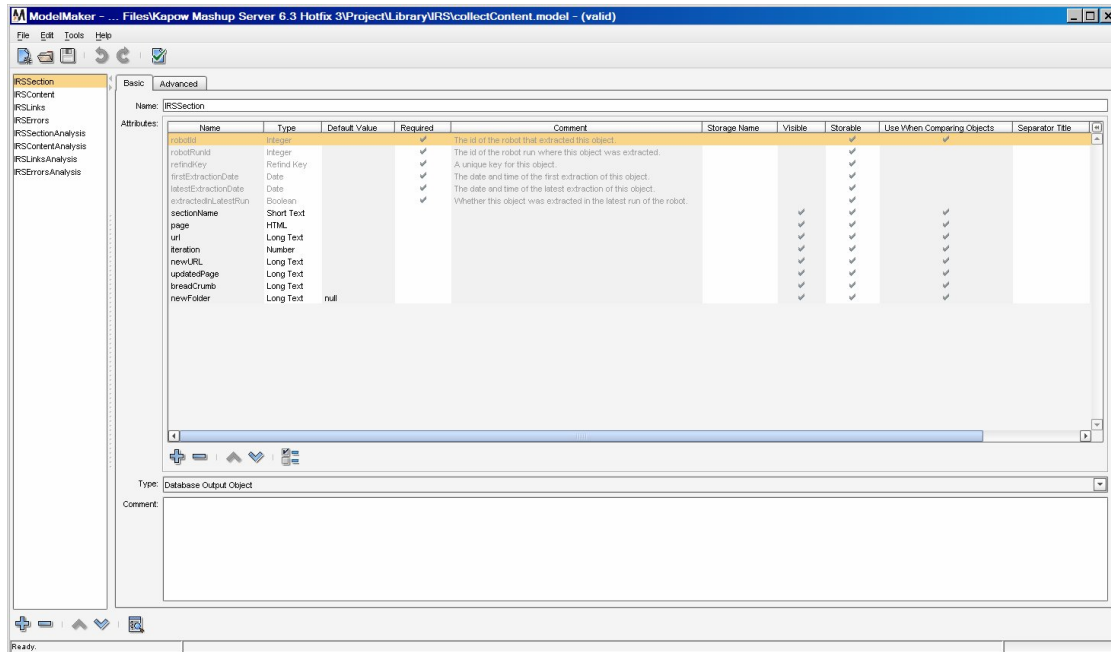


Figure 1: Content Model

Step 2: Content Collection

Once the content is modeled, the next step is to build the robot to collect content from the .gov domain. The robot viewed all the pages on the .gov domain, collected the content and stored the content in a database.

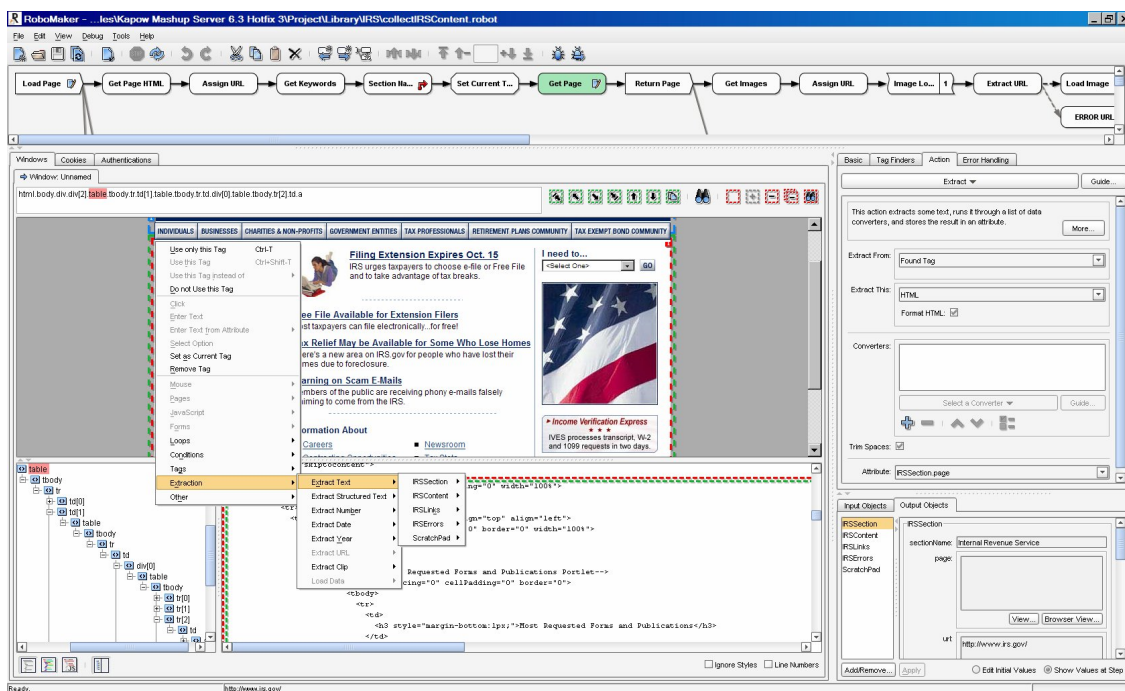


Figure 2: Content Collection Robot

NOTE: Our approach is very flexible, allowing you to collect the content through the CMS API's, CMS user interface or through the .gov website.

The flow of this robot is as follows:

- Load the .gov web site.
- Crawl the .gov domain
- On each page
 - Collect metadata on the web page [title, creation date, URL, author etc.]
 - Collect HTML content sections and remove headers and footers.
 - Loop over all images and store them.
 - Loop over all documents [PDF, PowerPoint, Excel etc.] and store them.
- Write the content into a staging database

Kapow supports all major commercial databases and for the purposes of this demonstration, we used MySQL. Part of the process is to build in error handling. Creating robots for production environments usually can be done in a matter of days, even for large scale migration projects.

Step 3: Content Transformation

The next step is to cleanse the data, performing transformation of content. For example, converting dates on web pages into standard database date formats; removing HTML tags, text and extra spaces from content.

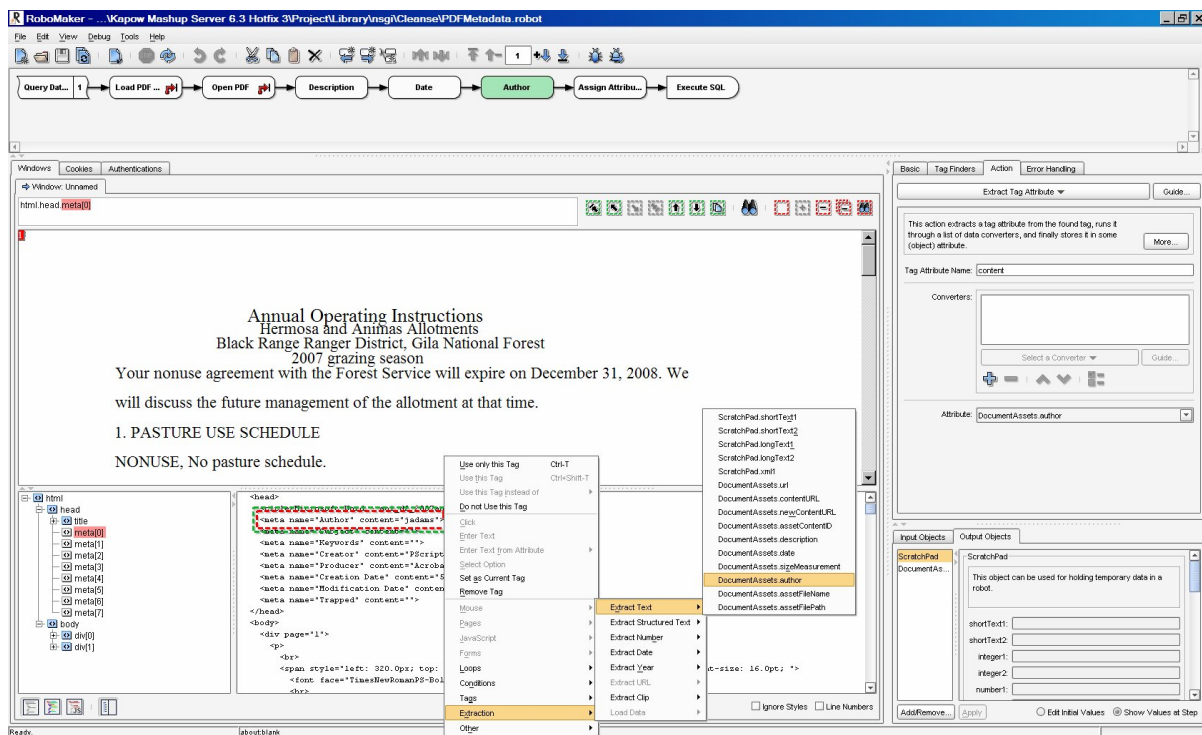


Figure 3: Content Transformation Robot

Step 4: Content Loading

After transforming the content, we can load it into the new CMS. The process of loading the content is essentially the same for any content management solution. The options for loading the content are standards-based so it works with all CMS systems. We provide three options: using the CMS content loading forms, working through the API, or loading directly into a database.

The next step is to take the cleansed content that is stored in the staging database and on the file system and load the content through the new CMS user interface or CMS API. Here are some of the capabilities Kapow provides for content loading.

- Content is loaded through the CMS API or the front end
 - All content is validated prior to loading.
 - Business rules of loading content are enforced
- Errors can be captured and either alternative action can be taken or they can be logged.
- Kapow can migrate or create new folder structures.
- When folder structures don't exist, Kapow can map existing structures such as 'bread-crumbs' into folder structures.
- Kapow can reformat content. For example, if an old CMS has e-mail content types and the new CMS does not, Kapow can convert the e-mails into text files and store them in an 'e-mail' folder within the new CMS.
- Kapow automates the process of re-linking content. Re-linking content is necessary because all content items are given new links when they are stored in the new CMS.

Step 5: Testing, Debugging and Deployment

After we've created the robot, we can test the robot in the debugging environment. The debugger is fully integrated into the development tool so that when an error occurs, it takes you to the point where the error occurred. Because the environment is visual, you see exactly what the robot was doing when it failed.

The screenshot displays the Kapow Integrated Debugger interface. The top window shows a flowchart of the robot's execution process, including steps like 'Return Object', 'Get Images', 'Assign URL', 'Image...', 'Extract URL', 'Load Object', 'Set PicFile Name', 'Get image ext...', 'Public Path', 'Write File', 'Get Iteration', and 'Return Binary...'. Below this, a table displays the output of the robot's execution, showing details for 'Internal Revenue Service' and 'IRS.gov Accessibility' content items. The bottom window shows a detailed view of the selected content item, including its HTML structure and metadata.

IRSSection	#	sectionName	page	url	iteration	newURL	updatePage	breadCrumb	newFolder
IRSSection	1	Internal Revenue Service	<table border="0" cellspacing="0" cellpadding="0" width="100%"> <tbody> <tr> <td class="...>	http://www.irs.gov/					null
IRSSection	2	IRS.gov Accessibility	<div class="contentPanel" id="contentPanel"> 	http://www.irs.gov/...	1				null
IRSSection	3	Internal Revenue Service	<div class="featuredContent">	http://www.irs.gov/...	2				null

Figure 4: Kapow Integrated Debugger
May 2008

Results

In a few days we were able to build a robot, run the robot against the site and collect a large subset of the .gov site. We also analyzed the site and found that most pages follow the same format.

Collected Content

- 2844 Web Pages
- 5892 PDF Files
- 688 GIF Files
- 23 JPG Files

We only collected a subset as an example but found the .gov site is consistently formatted. This further increases the value that Kapow brings to this Content Migration project because the consistency in the site design provides tremendous productivity and scale in automating the collection of content for migration to the target CMS and significantly accelerates the already rapid process of robot building.

Other Customer Examples

U.S. Department of Defense

A large DOD client chose Kapow for their content migration 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 tradition CMS capabilities.

Problem:

- Time Limitations: their current software was expiring
- Content Incongruity: they had content types in source systems that were not represented in the new system. For example, e-mails existed in one system but not within the other.
- Multi-level folder structures had to be retained
- Manual migration was too costly and time consuming to consider

Results:

- Kapow software enabled the client to migrate the content in 1/10 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 target system's calendar.
- Multi-level folder hierarchies were migrated.

U.S. Health and Human Services

The U.S. Department of Health and Human Services had multiple public facing web sites and portals. They wanted to migrate the content from multiple web sites and CMS into a single CMS but it was too time consuming and costly using traditional methods. They also wanted to know analyze the central CMS to identify issues with the content such as broken links. Finally, as they changed standards, such as including or not including abstracts to each content item, they wanted an automated way to make the changes across all of the content in the CMS.

Problem:

- Migrating content by hand was taking too much time and money.
- They needed to know which links to content were broken.
- They needed to clean up content in an automated manner.
- The HHS web presence is spread across many separate portals, web sites and CMS making it impossible to see all the information in one place.
- Each site has its own unique look and feel creating in inconsistent messaging.
- Source applications were written in different web technologies.

Results:

- Content was migrated in a fraction of the time and cost of their old process.
- Content analysis such as identifying broken links was automated.
- Content updates such as making every piece of contents' "abstract" sections not viewable were automated.
- Web assets and portals were turned into JSR-168 portlets and brought into a single portal providing a central location to view all HHS assets.
- Web assets brought into the portal were transformed to take on a common look and feel.
- A single search mechanism performs a federated search across all HHS portlets and search result links are clipped into a portlet with the same look and feel of the HHS central site so that links do not take users off the central site.
- 90% reduction of resources and time.
- No UI redesign was needed.

For more information about Kapow, visit www.kapowtech.com

For additional details about openkapow, visit openkapow.com.

Conclusion

Kapow's software automates the entire content migration process resulting in a solution that is dramatically faster, more reliable and less costly. By automating the migration process we are able to significantly decrease migration time as well as improve accuracy, removing the need to code complex business rules and content validation. Kapow is the best solution for Content Migration and we will prove it with a custom demo in your environment at no charge.

Kapow's Content Migration Edition product creates an automated, easily repeatable process similar to writing custom code, but leverages existing CMS user interfaces and API's. Our robot-based approach is significantly faster, easier to maintain, manage and adapt to changing business rules and migration strategies.

By automating the process of loading content through the user interface and/or the API, clients achieve these benefits:

- Robot creation takes hours and days instead of weeks and months because of the productivity increase provided by our visual, wizard driven, point-and-click development environment
- Improve accuracy through elimination of manual migration steps
- View, add and update custom content just like other content
- Content Re-linking; HTML pages contain links to other content. When the HTML pages and binary content are loaded into the new CMS they are given new link addresses, automating the management and updating of the links.

These are the additional advantages of using Kapow Software for content migration.

- **Content Analysis;** Kapow can analyze content before migration begins.
- **Partial Migration;** Kapow can migrate an entire site, a subpart of the site or a single record.
- **Incremental Migration;** Migration can selectively collect only new and updated content.
- **Multiple Sites;** Kapow can reuse robots that clean and load data, minimizing robot building
- **Duplicate Titles;** Kapow can eliminate duplicate data.
- **Security;** Security mappings can be migrated in the same manner as content.
- **Error Handling;** Kapow provides error handling, logging and notifications eliminating the need to write code, which needs to be tested and maintained.

Content Migration Cost Comparison

Content Migration Scope & Assumptions			
Task Description	Items		Labor Cost
HTML pages to migrate	10,000		
Binary Content (avg 2 binary links/page)	20,000		
Labor cost /hour; in-house (burdened)			\$ 65
Labor cost /hour; Kapow Prof Svc's			\$200

Kapow Migration Approach (Content Migration Edition)			
	S/W Costs	Hours	Labor Cost
Typical 2 yr Software Subscription (2-CPU license)	\$120,000		
Software Maintenance (included w/subscription)	Incl.		
Robot Development		80	\$16,000
Items requiring human intervention (HTML pages binary content not handled by the robot)		5%	
Manual processing time for exceptions (Time is significantly reduced because the content that can't be handled is still cleaned before being placed in the exception bucket) estimated 10 minutes per item		250	\$16,250
Subtotal	\$120,000	330	\$32,250
Total Cost - Kapow Solution		330	\$ 152,250

Manual Migration Approach			
Migration Time	Minutes	Hours	Labor Cost
Page Migration (per-page)	25		
Binary Item Migration (ea.)	10		
Page Content Re-linking (per-page)	10		
Total Page Migration (25 minutes x 10k pages)		4,166	\$270,833
Total Binary Item Migration (10 min X 20k items)		3,333	\$216,666
Page Content Re-linking (10 minutes x 10k pages)		1,666	\$108,290
Approximate number of errors assuming a 10% error rate (per page) [10k pages x 10%]	1,000 Errors		
Approximate time to fix errors assuming 10 minutes per error (this is the same time it would take for the automated process)		500	\$32,500
Total Cost - Manual Migration		9,665	\$ 628,289

Comparison Summary			
	Kapow	Manual	Savings
Migration Labor Hours	330	9,665	9,335 (96%)
Migration Cost	\$152,250	\$628,289	\$476,039 (76%)

About Kapow Technologies

Kapow Technologies is the expert in harvesting mashable public and private web intelligence with software that turns the Web into the world's largest database. Based on the company's patented, visual scripting method, the Kapow Mashup Server powers solutions in reputation management, competitive intelligence, asymmetric intelligence, content aggregation, content syndication, content migration, SOA enablement and business automation at a fraction of the time and cost compared to traditional software methods. As the number one solution in its class for three years running, the Kapow Mashup Server has attracted more than 250 clients such as AT&T, Bank of America, Wells Fargo, CSFB, Intel, Vodafone, Audi, DHL and SimplyHired, and partners such as BEA Systems, IBM and Oracle to power their solutions.

For more information about Kapow, visit www.kapowtech.com