monarch\\digital

Drupal Data Conversions: A White Paper

May 2015

Executive Summary

Drupal is one of the most popular open-sourced frameworks for content development and management. It's utilized by some of the biggest corporations and government institutions to share data and content over the Internet. But despite those merits it is still just one tool in a vast sea of solutions for storing and sharing content.

Being able to have your information move in and out of Drupal through other formats, and have your Drupal website able to facilitate other non-Drupal systems or databases is a valuable capability. We'll demonstrate here some of the methods we've employed for converting data through Drupal, and how you can leverage these functionalities to increase your website's capability.

Types of Information and Data Formats

There as many different ways to store and share data as there are speaking languages in the world. Making these varied formats accessible or useable through a single channel like a website is one of the most significant challenges in the business world. This is relevant as much for importing data into the website as it is for exporting it out of the website to be used in other systems or tools.

CSV or Excel spreadsheets

Spreadsheet files are a portable and easy to use format for reviewing data or evening migrating data between platforms. There are plenty of software solutions for working with spreadsheets (Microsoft Excel, Google Docs) which makes them a convenient solution for quickly creating reports. The standardization of the CSV file format affords it the capability of being a method for transferring large tables of data between systems.

Outside databases

Drupal is a database-driven content framework, and is most often coupled with the MySQL database server, but that isn't the only type of database software out there and most of the time the data contained in them isn't in a structure that Drupal can immediately consume.

Feeds

Feeds might encompass XML or RSS based content, and while typically employed for use cases such as syndicating content across a network of websites, can also be employed just as well for migrating that content into the website's database. Feeds like this are really just another way of encoding your website data in a way that other computer systems can understand.

Business Cases for Converting Data

Importing data and feeds

There are as many different reasons for importing data in your website as there are types of information you might be importing. A basic example would be a scenario where you need to import some additional data about customers you already have in your system, such as additional address information that wasn't requested when they originally registered at your website. You may have this information in spreadsheets, but want to have it added to the customer records in your website's database.

Mapping data for Drupal

Sometimes though, it isn't enough to just import some data to existing records in your website. It is also often the case that you need to take this outside data and restructure so that it can be used in totally new types of content in your Drupal website. Because Drupal works with data and database tables in very specific ways, simply importing the outside database tables won't make it immediately exposed to the various module that Drupal has for display content, such as the Views module.

Cleansing data

Sanitizing or cleansing data is a process by which you scramble or outright remove pieces of sensitive information in order to protect identities or prevent exposure of otherwise sensitive information. Email addresses, credit card numbers, and social security numbers are all excellent examples of types of information that will be targeted for sanitization from a database. This might be as simple as replacing the data with dummy values or blanks, and isn't really an attempt to encrypt the original information.

Data cleansing is an iterative set of processes that is centered around business rules and standards of acceptable data quality levels. These processes include investigative jobs to provide additional detail in detecting data patterns and data correction jobs to fill in missing or incomplete data and correct data values.

The situations where you might need to do this are varied, but will typically involve the cloning of a website or database for one reason or another. For example, perhaps you want to create a clone of your website for a developer team to make use of to create and test new features. You'll hardly want outside parties to have access to sensitive customer data, so cleansing that information from the database before handing it over to the developers is very important.

Feeding data to outside systems
It's also possible that you mean need to make some or all of your data available to outside systems that are already developed. Perhaps you had several apps that work in conjunction to provide functionality for your organization, and only part of that suite of tools was set up to work inside of a Drupal website. You may need to still expose some of your data to your other systems, so some conversion work will be necessary to extract that data out of your Drupal-based website and put it either into other databases or files that are consumable by the outside system.

Specific Use Case Examples

Custom Integration for MailChimp

The Business Enterprise Institute wanted MailChimp integration with their site, but the Drupal MailChimp module didn't provide enough of the necessary functionality on its own. BEI needed full list, subscriber, and campaign integration as well as a robust suite of statistics and charting. Using the Drupal MailChimp contributed module to provide access to the MailChimp API, we were able to enhance the basics of what the email system provided and facilitate control of their email campaigns through the Drupal website.

<u>Some of the extras that were added:</u>

- Dynamic campaign stats
- Dynamic subscriber stats
- Campaign creation on Drupal
- Campaign sending from Drupal
- Immediate send
- Scheduled send
- Creation and syncing of MailChimp lists.
- Custom webhooks on lists to update subscriber based on changes from MailChimp.

Building all of this functionality essentially provided the core MailChimp product into the Drupal site, but with more comprehensive statistics and finer control. The MailChimp UI can still be used if necessary, but isn't required. BEI can now control hundreds of lists and tens of thousands of subscribers from within their own website. Monarch Digital prides itself on clean functional coding standards, and the end results made us very happy.

Mass Product Price Update with Excel Files

Access Products prides itself on its customer service. The company has government contracts to provide printing supplies to a number of government agencies. For a number of business reasons, Access Products needed to keep its master product catalog in a database located at their site.

We developed an easy product upload and import process that uploads and updates the online product catalog. As all of the client organizations are using the same database, all customers receive the updates immediately and Access Products personnel only need to upload the updated product catalog once. The product import module utilizes the Drupal Batch API to facilitate a progressive update process.

Their staff are able to conveniently update thousands of products at a time, including the addition of product images and more than a dozen other product attributes. If products are found to already exist in the database they are updated, otherwise if not found a new product entry will be created. Images referenced in the import file will be looked for in the standard file storage area for the Drupal website, and the update process will ensure proper Drupal file management is handled, either storing the image references appropriately for the product or removing as well if required.

Drupal Best Practices for Converting Data

Batch Processing

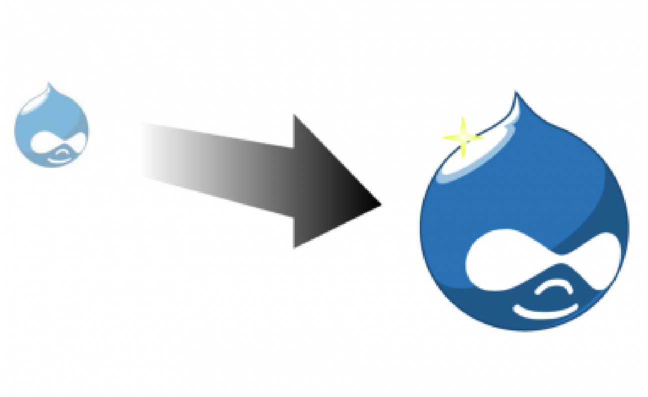
Depending on the size and type of website you are operating, the amount of resources required to export and convert its data can vary dramatically. Whether or not those resource requirements will become a problem if you run some kind of export will also be dependent on how powerful a web host you run your website.

But let’s assume for a minute that you are like most business operators out there and you don’t have unlimited funds available for operating a supercomputer to run your website. You’ll want to take care that any resource intensive operations are conducted in such a way as to minimize their impact.

With respect to running mass exports of data or conversion of tens or hundreds of thousands of records, you will want to organize these operations into batch processes. The Drupal Batch API provides the capability to collate jobs that your website needs to run into discrete operations, that can then be queued and run in small amounts at a time until all of the jobs are completed.

By letting these jobs run in small amounts, it prevents the system from "timing out" due to long execution time, it keeps the server from being locked up on one long process, and it provides breakpoints between each batch that give you a chance to get updates on the status of the job or log messages if things go wrong on a specific job.

Migrate Framework



Another robust framework of tools for converting Drupal website content is the Drupal Migrate module. This collection of modules is the culmination of thousands of manhours of work by the Drupal community.

It is more of a "framework" rather than just a "module, because it provides a toolset of reusable code by which a Drupal developer can map, convert, and migrate the content between different versions of Drupal websites, or even other systems like Wordpress or HTML websites into Drupal. In this sense it is not necessarily a complete out of the box solution as much as it is a toolbox by which you can develop a custom migration package for your specific needs.

There are a few reasons why the Migrate module and framework is so useful. Firstly, while it requires varying amounts of work on you or your developers part to set up the migration scripts, this system is the most high fidelity method for converting your content and Drupal website configuration to a newer version. Even more so, due to the continuing advancement of the Drupal system, a simple "upgrade" button is rarely sufficient for upgrading your Drupal website from one major version to the next. The more complex your website is the more this seems to be the case.

Additionally, the Migrate module hooks into the core Drupal management interface and provides you with a set of tools to monitor and even roll-back migrations of your website's content on a very granular basis. If something were to go wrong, and perhaps your user accounts migrate over just fine but maybe your order records do not, you don't have to start over from the beginning. You can evaluate the migration logs for the specific components that had problems, roll them back and make changes, and then try again.

Feeds and Export modules
If perhaps your needs do not require a wholesale migration or conversion of your entire Drupal website, but rather just specific types of content or pages, then the myriad of feed and export modules that the Drupal community has developed will likely be the right solution.

The Feeds module is useful for taking existing content in your Drupal website and exposing it via formats like RSS or XML so that it can be syndicated to other websites or software. While feed reader software may not be as popular among general web users today, the practice of syndicating content inside of a network of websites is still very much in vogue. Tools like this make it very easy to manage the creation and editing of content on one platform, and publish that content elsewhere where it will be read by visitors.

In other cases you may want to export some of your website's content into files on your local computer. These files could be CSV or Excel type of format so as to be used and manipulated in spreadsheet software. Or perhaps you want to export the content in order to transfer it to another version of the website for development purposes or to have a backup in the case of data loss.

There are huge variety of modules and tools in Drupal to facilitate these kinds of activities, such as the exportable nature of all of the modules plugged into the Drupal Ctools system, such as Views, Rules, Flags, etc. As well, modules designed around backing up your data like the Backup and Migrate module, make it easy to grab copies of your website database and save them for later restoration.

For collections of functionality, the Features and Node Export modules make it possible to combine not only the configurations that are used to organize your content, but also the content itself, so that you can reuse that work for other projects or deploy your website to a different web host in the case of a migration.

Data Conversion and Migration Pain Points

Lengthy Iterative Migrations
As mentioned before, data conversions and data cleansing is typically an iterative process. You can't realistically predict what kinds of edge cases you are going to run into, and many times your assumptions about how the data will be presented will be challenged.

Unfortunately, it is sometimes the case that these edge cases do not present themselves until running through a length migration or conversion process. The larger the dataset, the more chance that conversion or cleansing process could take hours or even days to complete, and each time you have to reset and debug, that timer will probably have to be restarted again.

When migrating into Drupal, sometimes trying to work exclusively inside of the Drupal API to accomplish can be one of the reasons for such a lengthy time for the migration or conversion process. The decision then becomes whether you can allow for more time to let the migration process run its course, or to budget for having developers create custom scripts to bypass the Drupal API and directly insert converted data into your system.

Every Situation is Different

These kinds of migration and cleansing projects don't lend themselves to letting you reuse a lot of your code. Of course, some of the underlying migration framework may be reusable or already provided, but the underlying source of the data or content for your conversion is almost guaranteed to be different every time. This will definitely make providing accurate estimates very challenging, as there are plenty of unknowns, seemingly similar projects will take very different trajectories.

The data cleansing process is probably the most difficult when the data you are working with is coming from user-provided input. For example, if you are trying to convert and migrate data from user profiles that have freeform input fields, and trying to match that input to very specific hierarchical taxonomies, the cleansing process can be quite laborious. It will be difficult to account for and guess all the ways that users might input their information in ways that don't match up to the assumptions provided by the taxonomy, and often the only way to discover these edge cases reliably is to just run through the conversion process and wait for it to run into problems.

Conclusion

Our intent for this white paper was to bring together the many concepts and techniques surrounding the process of converting or migrating data for Drupal websites. With the ever increasing number of software systems becoming necessary to effectively operate and manage your organization, you can be sure that a project like those will at some point become necessary. At Monarch Digital, we have extensive experience with conducting data cleansing, conversion, and website migration projects. Feel free to call or email.