



Hard Truths About Content Conversion

Joe Gollner

VP Enterprise Publishing Solutions
Stilo International

When organizations first look at migrating their content from unstructured to structured forms, there appears to be some common patterns that emerge.

Sometimes the people involved don't believe that adding structure to content is something that automation can provide any assistance with. We can call these people the "*skeptics*". Teams operating on this assumption budget time to perform the conversions manually - diligently adding markup to text files exported from their legacy authoring environment or, as a personal favorite, copying and pasting from the legacy tool into a structured editor. Sometimes these people happen upon a feature or two in their legacy authoring environment or in their chosen structured editor that helps to expedite the process in small ways. At the end of the day, however, this conversion pattern entails a massive amount of work and, most heartbreaking of all, the quality of the markup produced is often poor.

Way on the other side of the spectrum, we find quite a different mindset. These people tend to believe that there just has to be a packaged solution out there that will magically convert all of their content with the push of a button. We can call these people the "*believers*". Teams operating with this viewpoint will acquire a tool and run it against their legacy content. There are commercial tools that, as in one particular shameless example, claim to be "*universal*" converters (a claim that is true so long as almost meaningless structuring is going to be acceptable). Disappointment invariably follows as the outputs from these tools will almost always fall short of expectations and needs. The usual response to initial failures is to seek another conversion tool, one that will deliver more fully on what they believe should be an easy undertaking. Sometimes a finger is pointed at the erratic formatting of the legacy content and efforts are initiated to "clean-up" the content before the automated tool is run. It is somewhat ironic to find teams that start out as "*believers*" and then find themselves working just as hard as the "*skeptics*" on manual interventions.

Another pattern that comes up is based on the assertion that the simplest thing to do is to bundle up the legacy content and ship it overseas for offshore conversion. While this is a completely credible approach in particular cases, what is frequently underestimated is the level of management and quality assurance overhead that will be associated with this outsourcing effort. Also, if the conversion is challenging, perhaps dealing with erratic formatting or with a particularly complex and semantically ambitious target structure, this model can run into significant problems.

cont/d 2/...

2/...

Yet another pattern sees the Information Technology (IT) department leap in and declare that this is just another programming challenge that they were born to solve. Experience shows that this pattern is in fact the least successful of all possible approaches to content conversion. These programming efforts tend to be expensive, take a significant amount of time before showing any progress, and introduce sometimes comical quality problems in the results. The root of the problem, in this case, is that content conversion is unfamiliar territory for the vast majority of developers and the tools with which they are most commonly familiar often do not handle central issues with processing content at all well.

So the question that forces its way to the surface is this - is there a pattern for migrating content from unstructured to structured forms that has been proven to be consistently effective? The answer to this question, fortunately, is “yes”.

There are a number of components that we find in a genuinely effective conversion pattern. Firstly, there is a responsible reliance on automation which means that automation is used to the maximum extent possible without it being let loose blindly. Secondly, the type of automation that is deployed will include core capabilities that have been designed from the ground up to address the specific problems of converting content. Thirdly, a formal process governs the conversion activities and facilitates the efficient interaction of editors and subject matter experts (who actually understand the meaning of the content) with the conversion process so that the highest level of quality is achieved. Fourthly, automation is aggressively deployed to support content quality assurance with this encompassing structural validation, fidelity confirmation of the converted content with its source, testing of the converted content against planned downstream uses, and facilitating review activities by stakeholders. Finally, the entire conversion process is designed to permit adaptation to deal with the oddities that will inevitably come up and to provide opportunities for the process to be improved as experience builds.

While this may sound like overkill for many circumstances, it is a model that can be tailored to suit even the smallest project provided that team can get access to good automation and to a proven framework within which to conduct their conversion activities.

[Stilo International](#) has been building advanced conversion environments for over 20 years and during these years countless millions of pages have been converted by organizations around the world using Stilo's OmniMark technology. OmniMark is a specialized programming language and execution environment that has been designed to address the unique challenges associated with processing content and converting content in particular. While it is one of the best kept secrets of the content management industry, in fact the largest implementations of content management and publishing solutions tend, to this very day, to rely very heavily upon conversion processes built using OmniMark. Recognizing that many project teams really want the results, to see their content migrated into a structured form they can use, Stilo has worked to embody the core elements of the proven conversion pattern introduced above and to make this capability broadly available to projects large and small. It is for this reason that Stilo has released its newest offering [Stilo Migrate](#), an online service that will help organizations migrate their unstructured FrameMaker and Microsoft Word files into DITA.

See <http://www.stilo.com/migrate> and <http://migrate.stilo.com> for more information about Stilo Migrate. See <http://www.stilo.com/omnimark> for more background information on the OmniMark Content Processing Platform.