How Much Does Document Conversion Really Cost?—A Guide to Conversion Cost Variables

Don Bridges, Data Conversion Laboratory, Inc.

Whether to boost efficiency or to maintain compliance with industry standards, organizations are increasingly turning to XML specifications for their documentation needs. XML specifications include general DTDs like S1000D and DITA, and industry-specific DTDs like SPL (Life Sciences) and XBRL (Finance), to name a few.

But gauging just how much it will cost to convert your documents to XML is no simple task; a multitude of factors interact to determine the per-page price of any conversion project.

Complicating the matter are the various avenues you may pursue in order to get your documentation into XML format. For instance, how do you know when it is best to rewrite, or when automated conversion tools might be your best option?

It seems as though misconceptions regarding conversion costs have discouraged many from reaping the benefits of XML documentation. Too often, document conversion costs are wildly exaggerated, making conversion seem too expensive to be cost-effective. These exaggerated prices stonewall many worthwhile projects and do a disservice to all those who stand to benefit from more efficient, more functional content.

The misinformation regarding conversion costs runs both ways; it’s also not uncommon to find those who think that automated conversion tools are magic bullets that allow for perfect conversions to be performed in-house at the push of a button, and for only the cost of the software itself. This too is misleading.

In reality, documentation conversions are neither as costly nor as inexpensive as many people seem to believe. A $0-per-page conversion done with an automated conversion tool is little more than a mirage; even the best conversion tools necessitate considerable investments in other resources before they can yield useable conversion results. Fortunately, the document conversion that costs more than $50 a page is also largely mythical, appearing in only exceptional circumstances (such as high-security military projects).

The pervasiveness of these misconceptions has inspired this article, which is intended to bust the myths of fantastically expensive (or inexpensive) document conversion prices. This paper’s objective is to serve as a resource for commercial organizations that are planning an XML conversion or trying to determine whether documentation conversion may be a cost-effective option.

How much does conversion actually cost?

Document conversions can range from a few dollars to several hundred dollars a page, but the vast majority of commercial conversions cost the client no more than $3–5 per page. Note: The prices cited in this article are for conversions that are not export-controlled; that is, they apply to conversions of data that can leave its country of origin. This would apply to commercial materials and other materials without specific security considerations. For conversion projects that cannot be sent offshore, expect your per-page price to be at least double the per-page prices listed in this paper.

In this article, we will consider the following:

1. From what kind of source material are you converting?
2. What is your target format?
3. What type of document are you converting?
4. Does your conversion require the review of a content expert?
5. Do you require graphic conversion or content reauthoring?
6. When are automated conversion tools appropriate?
7. What other costs are associated with conversion?

Consideration #1: From what kind of source material are you converting?

As a rule, the more sophisticated the source format, the cheaper it will be to convert. Simpler source formats like paper and image-only PDF are the most expensive, since they require extra steps to extract text from the documents. On the other hand, source data in a more advanced format, like that found in documents produced by a word processor, does not require these extra steps and will be less expensive to convert.
Paper, Page Images, and Image-Only PDF These are the most expensive source formats to convert from because they require the additional production steps of optical character recognition (OCR) and proofreading.

PDF Normal These are the PDF files usually produced by word processing and publishing systems. Unlike image-only PDF files, which are just scanned images of pages, PDF Normal (also known as “searchable PDF”) files do contain the full text of the document. Since there is no need for OCR and the need for subsequent proofreading is largely eliminated, converting from PDF Normal costs less than converting from paper or images.

Word Processors and Publishing Systems In addition to containing all the text in a computer-readable form, documents created by word processors like Word or FrameMaker, or publishing systems like Interleaf or Quark, will also frequently contain styling and tagging information. If these have been applied consistently, they further reduce the cost of conversion.

It is much easier to convert from consistently styled documents since in many cases the styles can be directly mapped to XML tags. In many cases, conversion software can be used on consistently styled documents to directly produce the desired final output. On the other hand, inconsistently styled documents demand a much more involved analysis to interpret the various styles appropriately. In such cases, it is sometimes easier to ignore the styles altogether; other times it is worthwhile to try to salvage the existing styles with a pre-tagging process.

For some kinds of documents, conversion software can do a more accurate job of auto-tagging when provided with pre-tagging for unusual constructs that might be misinterpreted by the software. Pre-tagging is a process in which unique clues (typically special symbols) are added to the legacy content before conversion, and the conversion software leverages the clues to apply the desired XML tagging. Pre-tagging can be expected to increase per-page price by 10–25 percent but will significantly reduce the cost of cleanup later.

XML In some cases, converting from one XML format to another XML format can be done inexpensively—provided that two conditions are met:

♦ There is a large volume of source material to be converted. Volume is an important factor, since the process of mapping from one tagged format to another will require an investment in analysis and programming. This investment is fixed and independent of the size of the document set, so this type of conversion makes sense only if the conversion project is large enough to make this investment cost-effective. This may be

a good option for converting thousands of pages, but for a conversion project consisting of only a few hundred pages, you may be better off pursuing another conversion strategy.

♦ Your source format contains the information required by the target format. In cases where the tagging of the initial XML conversion did not capture information required by the target format, the missing information must then be retrieved from the source documents. The amount of information and the difficulty of obtaining that information from the source format will determine if converting from the tagged source is feasible or not. This is frequently an issue when converting from a structure-based DTD like DocBook to a content-based DTD like DITA, since the content-based DTDs require more information than the structure-based DTDs. In cases where much information has been lost, it might be better to go back to the original document.

For these reasons, for smaller projects as well as for those projects whose XML files are missing the information required by the target format, it may be more cost-effective to convert from your pre-XMl source data.

CONSIDERATION #2: WHAT IS YOUR TARGET FORMAT? The way that your target format organizes information also has a bearing on the per-page conversion cost. Converting to a simpler DTD that tags data by its appearance is cheaper, while converting to a more complex DTD that tags data according to function will cost more.

Structure-Based DTDs Converting to structure-based DTDs is relatively uncomplicated, since most of the chunks of information can be identified by their structure (such as a Section Header, Warning, Table, and so on). Therefore, the need for analysis, programming, and human involvement is reduced, and as a result, the overall cost per page will be lower. DocBook is an example of a structure-based specification.

Content-Based DTDs Converting to content-based DTDs is more complicated, since data chunks are tagged based on their content rather than on their structure. That is to say, where structure-based DTDs are concerned only with appearance, content-based DTDs are interested in substance. For example, when looked at from a structural perspective, a table is a simple arrangement of cells; however, a content-based DTD must look into the role played by the data within the table cells, which is a much more complex task.

Since the definition of tags is more complicated in content-based DTDs, sophisticated software is needed in order to recognize the tags associated with a particular chunk of data. It is therefore more expensive to convert to a content-based specification than to a structure-based specification. DITA and
S1000D are all examples of content-based tagging specifications.

**Consideration #3: What type of document are you converting?**

The nature of the documents being converted can affect per-page price as well. It will be easier and cheaper to convert a set of simple instruction manuals that are all similar to each other than it will be to convert a set of documents comprising multiple complex manual types. To the extent that each source or target manual type requires its own “mini-conversion” of any unique features, the more manual types involved, the more expensive the conversion will be.

**Number of Manual Types**

If the conversion is to be performed correctly, each manual type has to be tagged in its own way. This is because each manual type requires different information. Since each manual type undergoes a separate mini-conversion, there is a correlation between the number of manual types contained in the library and the overall cost of conversion.

An adjunct to this is the number of target DTDs to which documents are being converted. For those specifications which comprise multiple DTDs, each DTD requires its own mini-conversion as well; the more mini-conversions required, the higher the per-page cost will be.

**Type of Manuals**

Certain manual types are inherently more difficult to convert. For instance, troubleshooting and operation manuals are usually more complex than standard maintenance manuals. Troubleshooting manuals, for example, tend to include more elements that require intensive conversion effort, like layered graphics and flowcharts.

**Source Manual Conformance to Target Specification**

In many conversions, the original source manuals are structured differently than the target specification. Other times, the source document doesn’t even conform to its own purported DTD. In these situations, analysis is required to determine if all the information contained in the source document is needed in the target format, and if all the information that is needed in the target format is actually contained in the source documents. The greater the extent to which the source manual conforms to the target specification, the easier and less expensive the conversion will be.

**Consideration #4: Does your conversion require the review of a content expert?**

If you are converting to a content-based DTD (see Consideration #2) and your documentation set includes highly technical or subject-specific material, a review by experts in the field may be necessary in order to ensure that the content is correctly interpreted and tagged. Those performing quality assurance may also need to be familiar with the documentation subject so that they can notice and correct any errors that may have occurred during conversion.

The services of content experts are more expensive than the services of those with more general knowledge. As a result, conversions that require content expertise will cost more, and those that do not require any subject-specific expertise will cost less; however, the additional cost of content experts can often be greatly reduced by the use of specialized software tools, and techniques such as separating the content into portions that require expert review and portions that do not.

**Consideration #5: Do you require graphic conversion or content reauthoring?**

While the above four considerations (source format, target format, manual type, and content expertise requirements) influence conversion cost within a range of $3–4 per page at most, the next two variables have the potential to affect overall conversion cost to a far greater degree: whether or not your conversion requires reauthoring or graphic conversion could mean the difference between paying $5 per page and paying $50 per page.

**Graphic Conversion**

The conversion of raster graphics into vector graphics can constitute a very significant portion of conversion cost. Depending on the type and complexity of the graphic, the cost of graphic conversion can be as little as $0.10 or as much as $100 per image.

- **Raster to raster.** The simplest graphic conversion, leaving raster images in raster format costs an average of $0.50 per image. While raster graphics are not editable, a raster-to-raster conversion can provide the image in the same quality as the original document, making it a reasonable option for many image conversions.

- **Raster to vector.** These are the conversions that can cost hundreds of dollars per image. Vector graphics are more functional and editable than raster graphics, but their high price means that many programs are unable to convert their entire libraries. The question then becomes whether these added layers of functionality are really needed for every graphic, and if so, whether they’re needed immediately.

- **Vector to vector.** Even if your graphics are already in vector format, automated vector format conversions are not perfect and may require extensive cleanup of the new graphics to fix inconsistencies between the original and converted versions. However, this process is still more straightforward—and typically much less expensive—than raster-to-vector graphic conversions.
Aside from raster-versus-vector distinctions, the specific type of graphic in question also influences the cost of a graphic conversion. In general, block diagrams are the least expensive, line drawings are more complex and cost more, and schematics are among the most costly graphics to convert. Added levels of functionality (for example, wire tracing or indications of flow) can also raise the price tag.

One way that some programs have dealt with the high price of vector conversion is by converting only selected graphics. Sometimes the graphics singled out for conversion are those that will eventually need to be modified anyway, or else graphics are converted on a piecemeal basis, one-by-one as they require modifications. Others view vector conversion as simply too pricy to pursue at all, opting not to convert any graphics to vector format. It is worth noting that there is nothing that requires text and graphics to be converted at the same time, and nothing to prevent going back to convert images at a later date if a need for vector graphics emerges.

**Content Reauthoring**

The cost for reauthoring manuals can be the same as authoring a new manual—sometimes several hundreds of dollars per page. While reauthoring will produce data that is perfectly compliant to the DTD that you are using, justifying the cost of reauthoring when other options are available can be a formidable task.

Because of the cost, some take the approach of selective reauthoring, reauthoring only when and where it is absolutely necessary. Others may decide to postpone reauthoring until they have to undergo a major modification (for example, in the event of equipment upgrades).

When looking into a reauthoring solution, another cost and logistical factor that must be taken into account is that data that has been reauthored frequently has to be reapproved for distribution by the manufacturer and other regulatory agencies. This approval process can take time and the cost associated with reapproval can be significant. This is another factor that contributes to the popularity of selective reauthoring performed as needed.

**Consideration #6: When are automated conversion tools appropriate?**

Automated conversion software can be an attractive option for those looking to cut conversion costs. While these tools can be helpful when used in the right situation, there is a risk in overestimating what automation can do, and in underestimating the ancillary costs associated with a do-it-yourself conversion.

No conversion, not even one as straightforward as XML to XML, can be completely automated. Most off-the-shelf tools can be expected to yield an accuracy rate of 80–90 percent, and this number decreases for less-structured source formats.

While a 10 percent error rate may seem trivial, in a large-scale conversion, 10 percent may turn into a very expensive quality-assurance project. First, all 100,000 pages will likely need to be inspected, which at 30 seconds-per-page will require 833 hours. Then, if 10 percent of your 100,000-page document set contains errors and requires rigorous quality assurance or repair (say, six minutes a page), then you can expect to spend another 1,000 hours making manual corrections to 10,000 pages.

For this reason, conversion tools are most effective when they are customized to the specific needs of your conversion project. This will raise the accuracy rate of your conversion and reduce the resources that must be dedicated to quality assurance, but it often requires significant programming resources.

In some situations, performing an in-house conversion with the help of an automatic conversion tool is the most cost-effective option. However, if you are considering pursuing this avenue, it’s crucial to take into account expenses other than the software itself—namely, engineering, quality assurance, personnel training, and the opportunity cost of reassigning staff to tasks outside their area of expertise—so that hidden costs don’t take you by surprise.

**Consideration #7: What other costs are associated with conversion?**

If the above considerations are the variables that can raise or lower the overall cost of your conversion project, then the three items that follow are the constants—that is, costs associated with every conversion project.

**Quality Assurance**

As is the case whenever any modification is made to content, after a conversion, the owner of the content will have to undertake quality assurance processes to ensure the fidelity of the converted material. While quality assurance must be performed for any conversion, the price of this “constant” can still vary; the higher the accuracy rate of the initial conversion, the lower the cost of quality assurance. It is far less expensive to review correct documents than it is to identify errors and have them fixed.

**Infrastructure Development**

If the task of sustaining the data will be your responsibility, the cost of a content management system and an XML authoring and rendering environment should not be overlooked.
Training
In cases where the task of sustaining the data is your responsibility, the training cost to implement and sustain an XML publishing environment can be significant.

The Last Word
There is no one-size-fits-all price for document conversion; an oversimplification of the issues that determine conversion cost could just as easily land you “in over your head” as make a perfectly reasonable conversion seem out of reach. In almost all cases, a conversion of text documentation to an XML DTD should not cost more than $3–$5 a page; in most cases, it will cost less than that. Whether you are evaluating what kind of conversion to pursue or looking to justify your conversion budget, a thorough understanding of the factors that affect the cost of conversion will allow you to avoid common misconceptions and make better-informed decisions about converting your documentation.  

1 Systems, Inc., is a leading provider of outsourced billing, customer care, and print and mail solutions and services for the convergent broadband and Direct Broadcast Satellite (DBS) markets. Clients include Adelphia, AOL Time Warner, Brighthouse, Comcast, and Echostar.