OmegaT

OmegaT compatibility HowTo

This HOWTO describes OmegaT's compatibility with other software products.

General comments

Since it is standard procedure for professional translators to receive and deliver texts in digital form, users of OmegaT are naturally interested in its compatibility with other software products. This HOWTO aims to provide information in this area.

A general observation: "compatibility" is rarely a black and white, "yes" or "no" issue. Where the vendor of a software product claims that his product is compatible with another piece of software, this compatibility is seldom 100%. Conversely, where products are clearly not directly compatible, it is often possible to find procedures by which they can work together. The question which must be answered is whether these procedures are acceptable in terms of the result and the effort involved, and the answer is likely to vary from one user to another. In other words, "compatibility" is not only about products, but also about workflows.

Operating systems

OmegaT runs on any operating system on which a suitable version of the Java Runtime Environment (JRE) can be run. At the present time, this includes all versions of Microsoft Windows from Windows 98 onwards, Mac OS X, and most Linux distributions.

Source texts (files for translation)

Refer to the OmegaT user manual for an up-to-date list of all supported file formats. The list below is **not** an exhaustive list of all file formats supported by OmegaT, but is limited to those of particular interest to typical users.

- **Microsoft Office 2007-2010 (Word, Excel, PowerPoint 2007-2010); Office Open XML:**

  OmegaT is capable of handling the current Microsoft Office formats directly, without conversion to or from other formats.

  The Microsoft Office 2007-2010 file formats (also known as Office Open XML, and with the file extensions .docx, .xlsx and .pptx) is radically different to the Office 97-2003 formats which they supersede. In fact, their structure is very similar to that of OpenOffice.org files: they consists of a zip archive containing multiple files, and the files containing the text are based upon the XML standard. This makes it much easier for CAT tools to support it and edit it directly. The Microsoft Office 2007-2010 file formats can be processed directly in OmegaT; more information can be found here.

  For users of Microsoft Office who do not (yet) have the 2007 version, conversion between Office Open XML and Microsoft Office 98-2003 can be achieved using a free plug-in converter available here from Microsoft. This converter is available for Windows and Mac.

  Office Open XML can be converted directly to Open Document Format and back by means of the ODF converter, available here. This utility requires a version of MS Office (XP/2003/2007).

  Mac OS X users can convert directly from Office Open XML to OpenDocument Format and back in NeoOffice.

  Linux users can use this version of the standalone conversion utility to convert directly between Office Open XML and OpenDocument Format.

  Recent versions of OpenOffice.org are able to convert between Microsoft Office 2007 and Open Document Format.

- **Microsoft Office 97-2003 (Microsoft Word/Excel/Powerpoint 2003):**

  These file formats are proprietary, binary, and until recently were not publicly documented, all characteristics which make them extremely difficult to handle in a CAT tool.

  Some CAT tools solve this problem, at least for MS Word, by working within Word itself; most other CAT tools convert to RTF and back, often invisibly for the translator.

  OmegaT does not support these formats directly. Instead, they must first be converted manually to Office Open XML (MS Office 2007/2010) or Open Document format (LibreOffice or OpenOffice.org) before they can be handled in OmegaT. (OmegaT is far from unique in this respect: many other CAT tools convert source files to a different format, in some cases in the background without the translator being aware of the fact, before they are translated.)

- **RTF (Rich Text Format):**

  RTF is structurally quite different from the Microsoft Office 98-2003 file format, but with regard to OmegaT, the same applies: it is not supported directly and must first be converted to .docx (MS Office 2007/2010) or Open Document format (LibreOffice/OpenOffice.org).

- **Open Document Format: LibreOffice/OpenOffice.org/NeoOffice Writer/Calc/Impress**

  These file formats are the equivalent of Microsoft Office's formats. The Open Document Format is an international standard and has replaced the former proprietary (but open) Star Office file format (two different but very similar file formats). OmegaT has file filters for both Open Document Format and for the Star Office format which it replaced. These file filters are excellent, and the risk of damaging the formatting of a LibreOffice, OpenOffice.org, etc. file during translation in OmegaT is extremely low.

- **HTML; XHTML:**

  HTML and XHTML, its XML equivalent, are the most common file format for web pages. Again, OmegaT has excellent file filters for both. Like OpenOffice.org files, these file formats can be translated in OmegaT with very little risk of corruption. It is however worth customizing the filter settings for optimum results.

Translation memories

An international standard exists for translation memories: TMX, or Translation Memory eXchange. It has been widely adopted and is supported by almost all current CAT tools.

The TMX standard exists both in different **versions** and in different **levels**. The distinction is important for compatibility purposes. The standard is still undergoing development; this is what the different versions refer to. The levels refer to the formatting information that is contained in the TMX file:

- Level 1 TMX files contain no formatting information.
- Level 2 TMX files contain formatting information, but these files are typically only compatible when the same CAT tool is used. In other words, if an OmegaT user finds a 100% match in a Level 2 OmegaT TMX file, it can be accepted without requiring adaptation, but the same would not be true for a Level 2 TMX file produced by a different CAT tool (or vice-versa). This has repercussions for workflows in which CAT tool users (typically customers) expect to receive translation memory files and to have 100% matches inserted automatically.
- Level 3 TMX files contain formatting information in a form that can be read by other CAT tools. Support for Level 3 in CAT tools is rare.

Certain other CAT tools (such as TRADOS) are able to export different TMX files in different versions. OmegaT supports all current versions of TMX, but is likely to deliver better match results if the TMX file is version 1.4b.

Tools which support different levels of TMX files are in principle still compatible with each other. The formatting information contained in the higher levels will be meaningless to the "other" tool, but the textual information can still be viewed, fuzzy matches found, etc.

OmegaT uses the international TMX standard as its native translation memory format. Some CAT tools still employ dedicated proprietary translation memory formats, but virtually all support the import and export of TMX files. In practice, it is therefore possible for translators to deliver translation memories to customers and vice-versa, and for the recipient to use these files for immediate or future reference; if the files are to be used within an automated workflow, however, the constraint described above applies.

Further points to note regarding TMX files:

The TMX standard contains definitions of what characters are permissible. Not all CAT tools are equally strict in their observance of these definitions; consequently, some CAT tools are unable to open TMX files produced by certain other CAT tools directly. OmegaT is generally observant of the conditions and tolerant of other tools' failure to observe them; should problems arise here, however, they can generally be resolved fairly easily by a search & replace in a text editor of the illegal character in the TMX file.

TMX files are in the Unicode encoding, but may be UTF-8 or UTF-16. TMX files produced on Windows systems may begin with a byte-order mark (BOM). This differences do not generally lead to compatibility problems.

Compatibility problems may be caused by differences in the language codes employed. OmegaT supports language codes in the format "xx", "XX", "xx-YY" and "XX-YY", where xx or XX is the language, yy or YY the region. Strictly speaking, the ISO standard for language codes requires "xx-YY" (for example: "en-GB" for British English); although this variant is supported by OmegaT, the default convention offered by OmegaT is "XX-YY", e.g. "EN-GB". OmegaT is

tolerant when reading TMX files: it will accept files with en-GB, en-US, en, EN, etc. Not all CAT tools exhibit the same tolerance and some may not therefore display the expected matches if the language codes are not sufficiently compliant. This incompatibility can be resolved by searching for and replacing the relevant language codes in the TMX file in a suitable text editor. Another possible source of incompatibility are three-digit language codes, which are not supported by OmegaT at all. (This incidentally is a limitation of Java, not of OmegaT itself.)

Points regarding proprietary translation memory files:

The traditional Wordfast translation memory file format is of particular interest owing to its simplicity: it consists of a plain-text file with a translation unit (segment) on each line in which the source and target are separated by a tab. This format can be converted easily to the TMX format by third-party utilities such as Wf2TMX. Anaphraseus can also be used for this purpose.

Glossary files

OmegaT's glossary files are plain-text files in the format:

source term <tab> target term <tab> additional information

Some other CAT tools are able to import and export glossary files in this format, or in a similar plain-text format which can be produced from it very easily (for example by a search & replace operation in Microsoft Word).

OmegaT is also capable of reading glossaries in TBX, the industry-standard format for glossary files.

OmegaT cannot import or read glossary files in proprietary binary formats, such as Trados Multiterm.

Bilingual CAT-tool formats

Many CAT tools make use of an intermediate bilingual file format, i.e. a file containing both source and target language segments, and in some cases also the structure of the original file. Originally, these bilingual file formats may have been a by-product of the tool's architecture. They have however become an important phenomenon in translation workflows involving CAT tools, and they often present the greatest obstacle to compatibility between OmegaT and other CAT tools (or for that matter between CAT tools in general).

There are at least three reasons why a customer may request delivery of a translation in a particular bilingual file format (rather than simply delivery of the translated file and possibly also the translation memory):

1. Some CAT tools, notably TRADOS, are able to import a wide range of file formats, including desktop publishing formats, and to prepare them for translation in the tool concerned. The "prepared" form is typically the tool's bilingual file format. Without preparation in this way, the original file format may be accessible to the translator.

2. The translation stage is only part of the customer's workflow. The translation may for example be passed to a checker for editing. If the checker's changes are to be included in a translation memory repository maintained by the customer, the changes need to be made *before* the final documents are created. This can be done either within the CAT tool concerned, or in some cases in an external bilingual file format which the tool is capable of reading.

3. The customer wishes to receive a translation memory against which texts can be run in the future: in other words, to "pretranslate" future texts against an existing translation memory. In order for this process to be as automatic as possible, two conditions must in particular be met: firstly, the translation memory must contain formatting information (see above); and secondly, the segmentation rules applied to the text must be the same as those applied when the translation memory (or part of it) was produced. The easiest way for a customers to ensure that these two conditions are met is for them to pretranslate the text themselves before passing it to the translator (thereby defining how it is segmented), and to create the translation memory in the CAT tool of their choice after receiving the translated

intermediate bilingual file from the translator (thereby ensuring that the formatting information contained within the translation memory will be compatible with future projects).

Several of the bilingual file formats can be handled by OmegaT, and not necessarily with great effort. An understanding of the processes involved is however important. The individual bilingual file formats are described below.

XLIFF

XLIFF is the industry-standard bilingual file format. It is supported by several CAT tools, and in fact some CAT tools are effectively "designed around" the XLIFF standard: Heartsome and Swordfish are examples. Since it is a standard, an advantage of XLIFF is that file filters provided by one CAT tool vendor for conversion between a certain format and XLIFF (and, following completion of the translation workflow, back again) can in theory be used to prepare files in the format concerned for translation in any CAT tool capable of supporting XLIFF. In practice, working with the XLIFF workflow often requires the use of tools that are not very user-friendly.

OmegaT has rudimentary support for XLIFF, and a procedure for using XLIFF in OmegaT in conjunction with the Rainbow tools can be found here. The filters available are mainly for file formats peculiar to the IT industry rather than end-user files.

Trados "uncleaned RTF"

The Trados "uncleaned" RTF file format, often referred to simply as "uncleaned files", has for many years been the most common bilingual file format used in translation workflows. It owes its origins to the use of MS Word as an interface for the Trados CAT tool. In addition to Trados, however, several other CAT tools also support the "uncleaned RTF" format, notably Wordfast Classic.

Essentially, this format consists of an RTF file in which the source and target segments alternate. These segments are marked and delimited by special characters and MS Word formatting styles.

A script (for Windows only) and procedure was recently (2008) developed enabling OmegaT users to produce Trados uncleaned RTF files for delivery at the end of their translation stage. For details, see the "Exporting from OmegaT to uncleaned RTF" HOWTO.

Trados TTX

Trados TTX format is the counterpart to the "uncleaned RTF" format for Trados Tag Editor, which unlike Trados Workbench, does not work in direct combination with MS Word. TTX is an XML-based format. An OmegaT plugin is now available by means of which this format can be handled in OmegaT.

Wordfast TXML

Wordfast TXML is the native internal format of Wordfast's new Wordfast Professional (also known as Wordfast 6.0). As its name suggests, it is an XML-based format. It is supported by OmegaT.

Déjà Vu "External View"

An interesting feature of Déjà Vu DVX is its "External View" file format. This file format enables OmegaT users to deliver bilingual files to users of Déjà Vu DVX, who are then able to edit them further or to incorporate them into automated workflows. For details, see the Déjà Vu "External View" HOWTO.