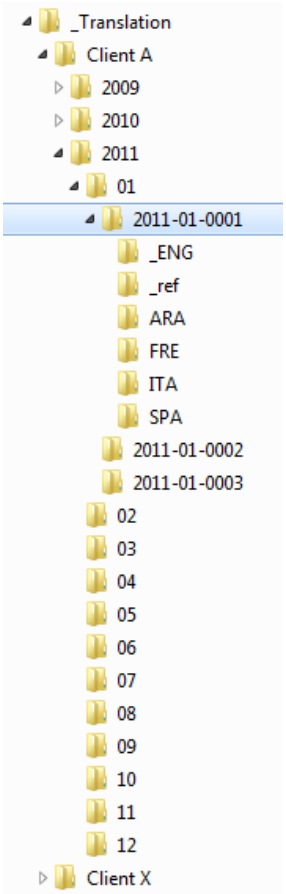
Good practices for archiving multilingual documents

At the first glance, multilingual document archiving may seem easy but there is more than meets the eye. Too many people fail miserably at naming files and categorizing them. In no way we pretend that this is the only method but we do hope that this document will give you useful insights on how to archive multilingual documents. This document is simply the cumulation of many years of experience that we deemed relevant to share with our clients. Methods exposed here are currently used by many organizations.

Creating a logical folder structure

The folder structure is the backbone of your archiving. What is the use of naming your files correctly if you don't know where they are? Each document is linked to a client, a domain, a project or anything else. This link is represented by a folder.

Here is an example of well-built folder structure:



The *_Translation Documents* folder, which is the root, is preceded with an underscore. In computing, the underscore is one of the characters that prevails on letters and numbers in an alphabetical sorting. That way, we can be certain that the folder will rank at the top of the folder structure, making the root easy to find. As far as the "numbered" folders are concerned, make sure you use the same number of digits for all of the folders. In the above example, months are written using a two-digit system (01, 02, 03, etc.). If you want to create up to 999 folders, you should think about using a three-digit system (001, 002, 003, etc.) instead of writing 1, 2, 3, etc., otherwise the folder logical sequence will fall apart.

- Under the root folder, each client has its folder (Client A, Client X, etc.).
- Under each client folder, we added year folders.
- Under each year folder, we added month folders represented by numbers.
- Under each month folder, we added project numbers. The project number should be relatively short while providing enough details about the request or the translation project like the year, month or anything that you see fit.
- Under each request number, we added language folders and the *_ref* folder. The latter is used for any reference materials pertaining to the project.

What interests us the most however is the structure of the language folders and the naming of the files. Let's have a look.

Let's take the structure on page 1. In the 2011-01-0001 folder, the "_ENG "folder is preceded with an undescore in order to identify the source language of the project. In this example we chose English as the source since English is more often than not the source language for most people. In a multilingual context, you need to identify the folder which contains the source documents with some sort of language marker otherwise translators will lose their time looking for source files. Time is money.

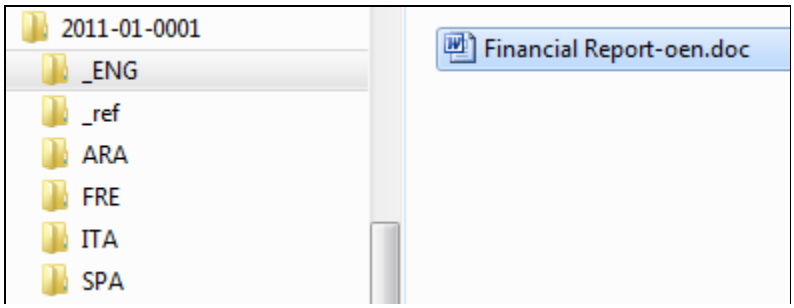
File Naming

Everybody has its way of naming files but certain methods have proved to be more effective than others. Here is a list of basic rules that should not be overlooked in file management, especially when translation work is involved:

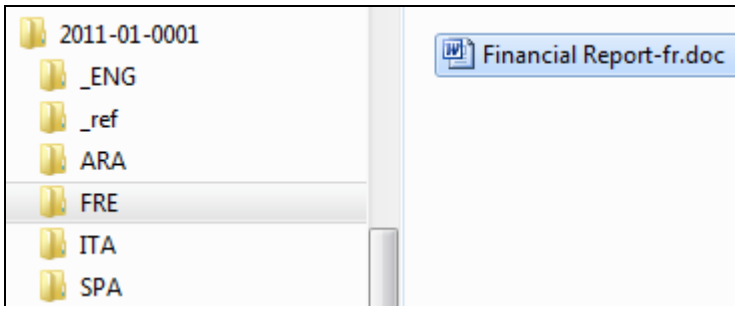
- Source and target documents should always bear the same name, or very similar name. Never translate the target files;
- Files should bear a language marker as a suffix not as a prefix otherwise matching files might not be close to each other in a long file list. The language marker should be separated from the filename itself using either the underscore or the dash. We do not suggest using the space as a separator;
- Filenames should be as short as possible while providing the most details about the content. If you have to open the document to know what is inside, it is inefficient.
- Avoid any accented character in your filenames and folders. Compression software (Winzip, 7zip, etc.) may refuse to zip or unzip files due to accented characters.

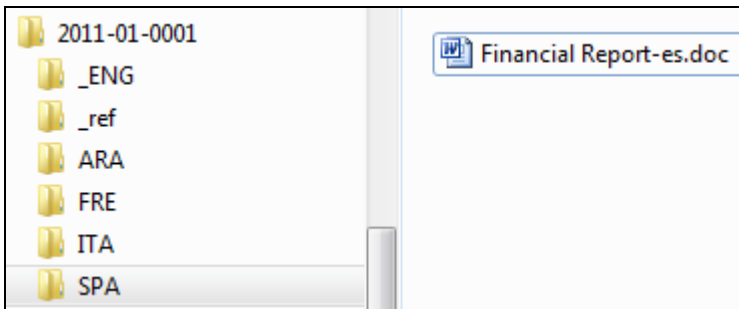
In a multilingual translation project, it can ben very useful to know what language is the source language just by looking at the filename instead of the folder. Languages markers are thus essential.

Here is an example:



In that example, the source language is English so we added the suffix “oen” meaning “original English” to the filename. Target-language documents will obviously not bear the “o” marker but only their regular language marker. Here are examples of target-language documents:

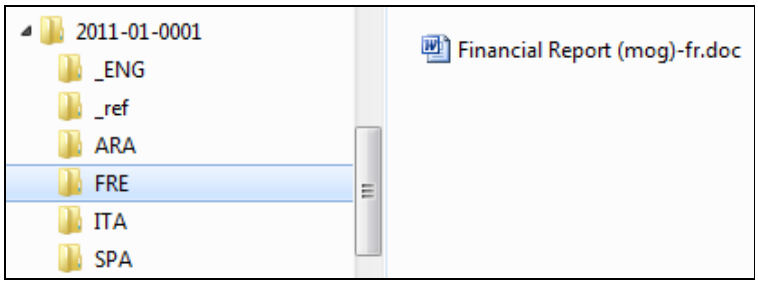


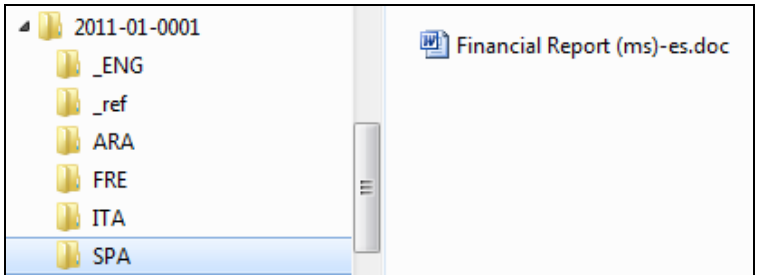


The choice of language markers are arbitrary. In the example above, we chose a two-letter system: French=fr, Arabic=ar, Spanish=es, Italian=it, English=en. Prior translating a project, a copy of the source document is placed in every target-language folder and renamed accordingly using the target-language markers. Translators may work in those folders directly.

Initials in filenames

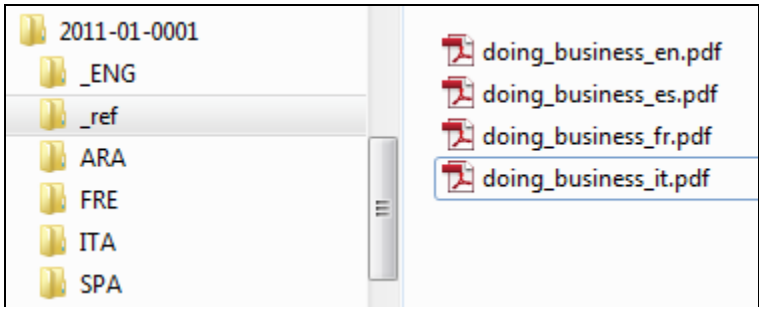
Adding initials in filenames give more details about who worked on the document. In the examples below, letters in parenthese are translators' or revisors' initials. Once again, we recommend inserting them towards the end of the filename so to not influence the alphabetical sorting.





Reference materials

For certain projects, you may be lucky enough to get reference materials of any kind (glossaries, notes, translated documents, etc.). In the following example, the client has provided past translations related to the new document to be translated. We renamed the four documents in order to be consistent with our naming convention.



To use these reference materials effectively, we suggest that you create bitexts with them (see next paragraph).

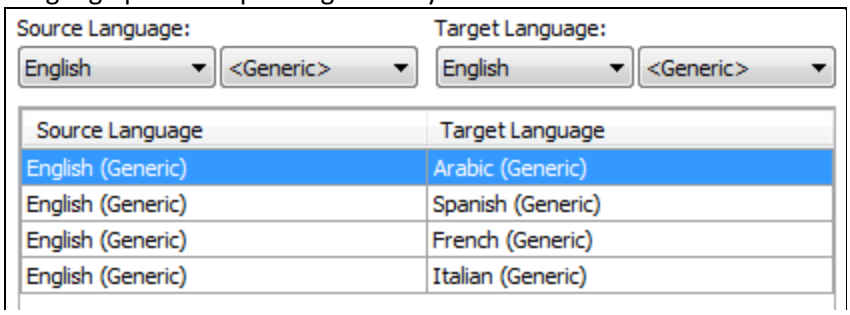
Creating bitexts from multilingual documents

Once the translation is done, you need to feed your translation memory. Depending on the tool you own, you may have to align documents prior importing them into your memory. If that is the case, here is a way to do it. We chose the tool AlignFactory to generate bitexts. A bitext a the merge of two documents shown side-by-side.

AlignFactory enables you to automate the pairing and alignment processes, unlike AlignFactory Light (manual pairing only). It is important to differentiate the pairing process from the alignment process. Examples shown here are based on the “automated pairing” feature and not on the “manual pairing” feature of AlignFactory.

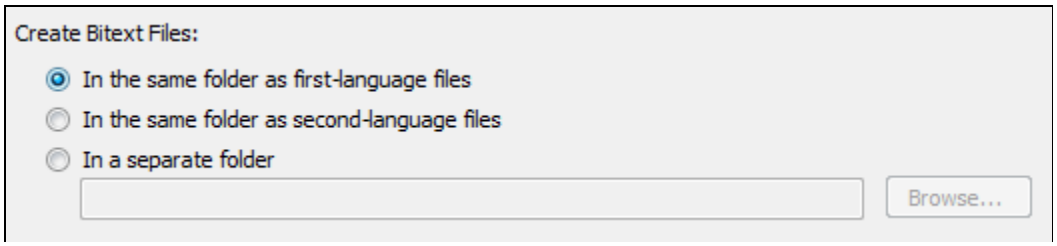
The pairing process is based on filenames which means that naming your files consistently and logically is of upmost importance. Before using AlignFactory, you need to set up language pairs that you will be using. You cannot create two pairs with the same languages even though the direction differs, for example, English-French and French-English. Since bitexts are bidirectional, this would be useless. For the following examples, we created four language pairs in which English comes first but that doesn’t mean it will always be the source language. The bitext name always contains the language pair combinaison. In the following example, it only means that English is the first column of the bitext and the other language is displayed in the second column: *filename*_ENG-FRA_BT.xml, *filename*_ENG-ARA_BT.xml, etc.

Language pairs setup in AlignFactory.

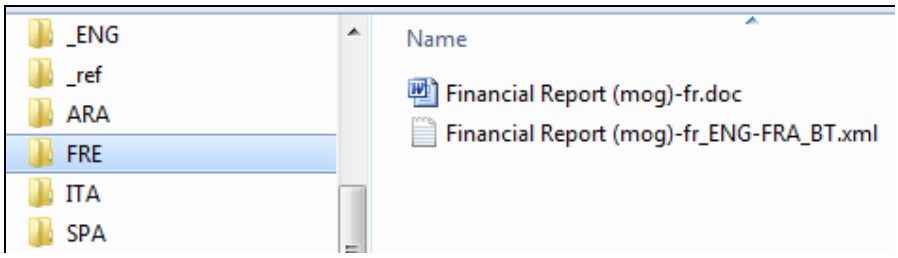


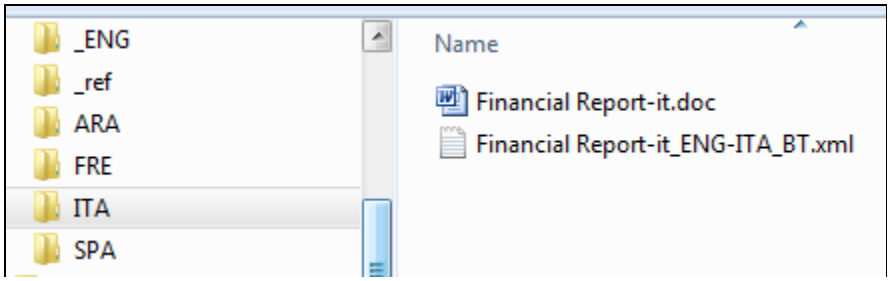
Once language pairs are set up, it is useless to swap languages should the source language of the project be different from English for a different project. This is why language markers are very important. You will see why in a moment.

In AlignFactory options, you can choose to create the bitext in the source- or the target-language folder. Should you choose to create the bitext in the source-language folder, the bitext will bear the source document name, otherwise it will bear the name of the target-language document.

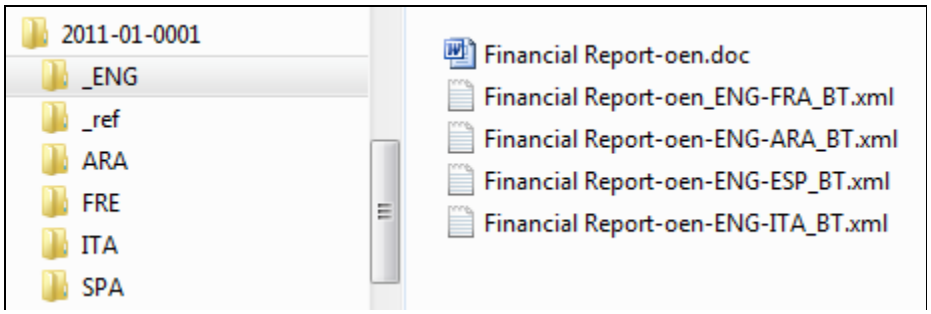


Here are two examples where the bitext has been created in the target-language folder.





The language pair in the bitext name is not as reliable as the filename itself. This is why we strongly recommand using language markers. In the two examples above, the marker « o » is nowhere to be seen in the filename. That means neither French nor Italian is the source language but since English figures in the language pair in the bitext name we can only come to the conclusion that English is the source language. Here is the opposite scenario where bitexts have been created in the source-language folder.



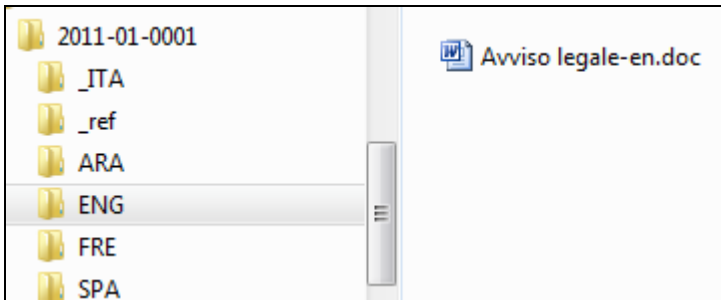
The “oen” marker proves that English is the source language.

English a as a target language

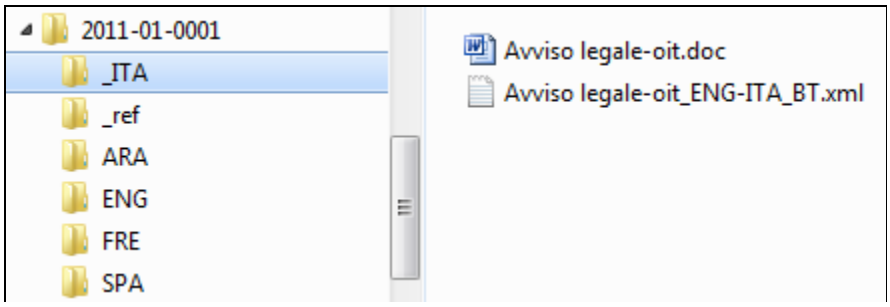
Let's use the same language pairs as the previous example. Suppose that the source language is Italian. Prior translating the document, we add the suffix "oit" to the filename meaning "original Italian". Notice that the Italian folder is preceded with an undescore.

Example:

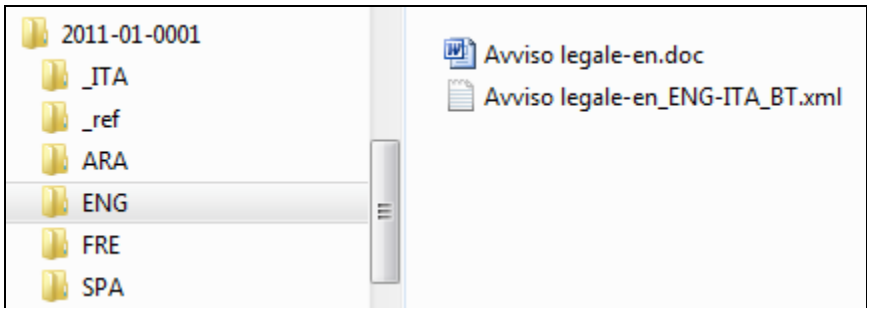
If you translate to English, you would make a copy of the source document in the English folder and name it this way:



Once the translation is done, you can create a bitext. If the language pair set in AlignFactory is ENG-ITA, and you choose to create the bitext in the target-language folder (in this case Italian is the target language because it is placed second in the language pair) the bitext will be named this way:

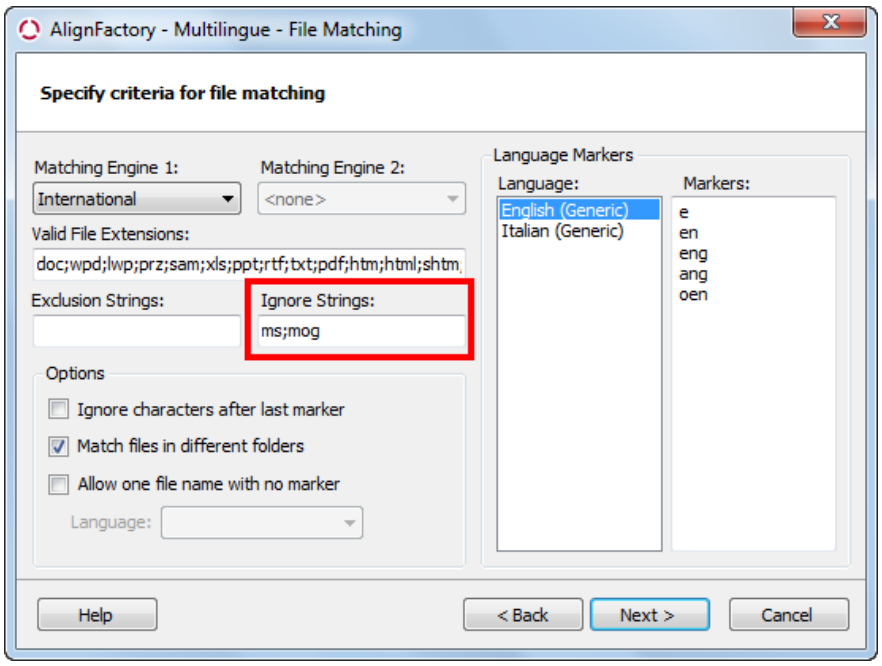


If you create the bitext in the source-language folder (*English* according to the language pair set in AlignFactory), here is how the file will be named:



Initials in filenames

If you add initials in filenames, you need to tell AlignFactory to ignore these so it can pair documents automatically prior aligning them. The **Ignore Strings** field enables you to neutralize anything in the filename.



Conclusion

Examples presented in this guide are only suggestions. We do not suppose that this the only method as each client has different needs. The main objective of this guide is to give guidelines from which you can base yourself on. Whether you possess a computer-aided translation tool or not, this method remains very relevant in any working environment since part of the efficiency of each employee depends on it. Should you have any suggestions or would like to share your method with us, please write to sales@terminotix.com.

The Terminotix Team