Term Extraction Starter Guide

by TerminOrgs

Version and date	Reason for release
v1 – 2022-03-29	Initial publication

Table of Contents

About this guide

Developed by: Terminology for Large Organizations (TerminOrgs)
Established in 2011, Terminology for Large Organizations (TerminOrgs) is a consortium of terminologists and other communications professionals who promote terminology management as an essential communications strategy in large organizations. This group is a forum to discuss and develop guidelines and best practices for large-scale terminology management. Our mission is to raise awareness about the role of terminology for effective communications, knowledge transfer, education, risk mitigation, content management, translation and global marketing, with a focus on large organizations. TerminOrgs represents stakeholders of terminology standards and tools. We work to determine and promote the economic value of managing terminology. To learn more about TerminOrgs and our activities, please visit www.terminorgs.net.

Document editor: Christine Hug, March 2022. This document may be updated periodically. Please visit http://www.terminorgs.net to download the latest version.

Who should read this guide?

People interested in using term extraction to improve their terminology management. We assume that you have already read the Terminology Starter Guide. (http://www.terminorgs.net/Terminology-Starter-Guide.html)
If you need a grounding in terminology, please consult the Pavel Terminology Tutorial (in English, French, Spanish, Portuguese, Italian, Arabic and Dutch at http://www.crtl.ca/Pavel/Pavel%20Terminology/www.bt-tb.tpsgc-pwgsc.gc.ca/btb66a0.html?lang=eng&cont=308) or DTT publications on best practices for terminology work or term extraction (http://dttev.org/dtt-publikationen.html German landing page, but some material available in English).

What is term extraction?

Term extraction (sometimes called term harvesting, or scanning for terms) is the selection of terms in a text or group of texts for later terminological study. It used to be done by terminologists who would take a paper page of text and use a highlighter to mark terms of interest. Once done, they would look at those highlighted terms and add relevant terms to their termbase by entering them manually. Now, there are software tools that can help automate or partially automate the task for scanning electronic documents for terms. We refer to them here as term extraction tools (or TE tools).

Monolingual term extraction extracts candidate terms from a text in one language (useful, for instance, for collecting the terms you need to translate and also for checking to see whether they are already in your database before you start translating texts). If you only write or communicate in one language, then monolingual is all you need.

Bilingual term extraction performs the monolingual extraction in the source text, but then looks in the matching target text segments in an aligned bilingual corpus to suggest terms that might be the equivalent, though results of automatic extractions need to be validated by a language expert. If you work in a multilingual environment, bilingual extraction will likely be a useful part of your terminology management. Bilingual term extraction processes bitexts,

bilingual corpora or translation memories (pairs of source and target texts, aligned by sentence or paragraph) and extracts terms from the source texts. Once these extractors search the target text segments matching those in which occurrences of a given source text candidate term are found, they then see which target text candidate terms (word combinations) appear most often in the matching target language segments. They either extract the best ones (highest probability of being equivalent) or require the user to select from a list of possible equivalent term candidates. Simplistically, imagine an aligned corpus as two sides of a cardigan buttoned together at equivalent intervals, with a source text candidate term in the source text segment to the left of buttons 1, 3 and 5. The tool would look for an equivalent term in the target text segments to the right of buttons 1, 3 and 5, checking for what they have in common.

Do I need a tool for just one language (monolingual)? Or a tool that can extract terms from one language and find the equivalents for those terms in translations or comparable texts in another language (bilingual)?

If your organization deals with only one language (no translation or transcreation), then monolingual extractors can help you improve internal and external communication (explaining acronyms, deciding which synonym is preferred or forbidden, for example).

If your organization needs to generate similar texts in multiple languages (the same user manual content in multiple languages, for example), then bilingual term extraction will likely help you mine bilingual documents for terms and their equivalents.

What are the different ways to perform term extraction, and how do I choose which to use?

Manual term extraction can be as basic as using a highlighter to mark terms on a paper document and transcribing them into your termbase. Or highlighting terms in a document electronically and copying/pasting them (and perhaps a context sentence) into the terminology record in the termbase.

Choose this option for very short documents or documents only available on paper. Sometimes, manual term extraction of a short paper document is faster than taking the time to scan and convert it. Also, some documents are too short to generate good results from a TE tool.

Semi-automated (or hybrid) term extraction means having the option of using the extraction tool to generate a list of term candidates, but then having a language professional go through this list to examine occurrences of the term candidate in context (with a corpus search or Key Word In Context tool), copying over term candidates worthy of being retained, potentially along with the sentences that contain them if they are useful. For example, if you see a term in context and the context sentence is a definition, then copying that definition into the definition field of that concept's term record is helpful. As you read such context sentences or read through the aligned corpus, you may come across other relevant terms that the tool overlooked and left off its list of term candidates. You can take that opportunity to record those manually.

Choose this option if your TE tool is statistical and does not automatically extract term candidates that appear only once in a corpus (minimum frequency 2), if your project requires that you also extract terms that appear only once. Also choose this option if you want your language professional to validate candidate terms earlier rather than later in the process.

Automatic term extraction means setting parameters for the extractor (for example: minimum frequency, maximum length of term, particular parts of speech only) and letting it generate a list of candidate terms in a source language, and maybe even propose an equivalent term from the matching equivalent text segments if the extraction is bilingual. Automatic term extraction tools process a monolingual corpus according to either statistical rules (for example, two words occurring together many times may be a term) or linguistic rules to extract term candidates without human intervention (except to set initial parameters), generating a list for later validation by a language professional. Most TE tools can even prepare draft records documenting a term candidate and the best-guess equivalent (when there are several possibilities). Here, apart from setting the parameters, intervention from the language professional is mainly at the end of the process, when the language professional validates (and completes, if necessary) the records for good term candidates and adds them to the termbase, while records for bad or irrelevant ones are either researched and corrected or deleted without ever being added to the termbase. The language professional can also examine the lists of candidate terms or the draft records to find and mark undesired synonyms (as not recommended, deprecated or forbidden, for example) in monolingual termbases and in bilingual/multilingual termbases to enable efficient use of the terminology data for controlled authoring and consistent translations. Automatic term extraction is offered by most tools.

Choose this option if your deadlines are extremely short and you have a corpus too large to scan manually or semi-automatically.

Note: Ideally, you select the term extraction approach and type of TE tool according to each project (deadline, available corpus, purpose, etc.), because while having just one approach and tool is better than none, one is not always appropriate for all projects.

What is it?
- Monolingual vs bilingual
- Manual, semi-automated or automated with post-validation

When should TE be carried out?

- Populating a termbase initially
- Preparing a glossary for a translation project or for internal communication
- Determining preferred terms or commonly used terms within a given language or set of texts and deciding on allowed or not-recommended synonyms, to enable both consistent monolingual text production (controlled authoring) and consistent translation.

- Checking coverage and completeness of your termbase
- Improving your termbase quality
- Preparing for terminology standardization committee work (a group discussing and coming to consensus about which terms to use for a given set of concepts)

When?
- Feed termbase
- Create translation glossary
- Ensure consistent internal/external communication
- Examine existing term use
- Select approved terms

Different types of TE tools

- **Statistical** term extraction tools can extract all words and look at the frequency of source text terms. These work at the same level of quality for most languages.
- **Linguistic or rules-based** term extraction (also called morphosyntactic) parses the text by part of speech and then looks for combinations, such as noun + noun, that are likely to be terms. Linguistic rules have to be made for each language the tool handles, so while they have some success identifying segments (word combinations) that may be terms, they only work for the languages for which the tool has the linguistic rules.
- **Hybrid** term extraction uses a combination of both approaches.

Types?
- Statistical
- Rules-based
- Hybrid

Which functions do TE tools offer?

(Note: No tool offers them all.)

- Extraction of both unigrams and multigrams (single-word terms and multiword terms). Some tools can do both at once, while others require a separate pass for each.
- Indicating the frequency with which the term appears in the corpus
- Use of an exclusion list or stopword list, which may be user-editable. A list of unigrams and multigrams not to include in the term candidate result list
- Use of a recognition list - marking/annotating the term to indicate it's in a termbase. This functionality helps with cross-verifying new data versus what is already in the

database. Some tools even verify against a second set of data (for example, against a publicly available glossary).

- A way to choose not to extract terms already in the database (delta extraction)
- Use of reference corpora - such tools compare the frequencies of term candidates found in the analysis corpus (texts with terms to be extracted) with those of terms extracted from a general language reference corpus. This helps term candidates for technical or specialized terms stand out from general language vocabulary.
- Use of a lemmatizing function to get the canonical forms from plurals, conjugated forms, etc.
- A way to group similar terms for examination
- Providing context(s) – to see occurrences of terms in context and allow contexts to be extracted with the term candidates.
- Indication of the name of the file the term/context was extracted from
- Marking of a term's part of speech
- Extracting acronyms matched to their full forms
- An option to extract proper nouns separately
- A filter by part of speech (e.g., extract only nouns or verbs or adjectives)
- Ability to search a variety of file types (PowerPoint, Excel, PDF, Word)
- Method for exporting an extraction list and preliminary bilingual term records in various formats, such as TXT, CSV, or XML
- Proposing equivalent terms in bilingual term extraction
- User-modifiable settings for minimum occurrence limit and term length
- Alignment of document pairs if you don't already have bitexts (source and target texts merged into one file with segments aligned) or TMX files (bilingual translation memories)

Which terms should you extract? How do you know which to choose?

Issue to consider	**Points to ponder**
Define termhood	First, define termhood (what is a term to us?) for your organization's situation and needs. Think about who will be needing to use these terms and why. Are you trying to improve communication internally only, or externally too? If the goal is standardizing to help a sales team communicate without making false claims that could lead to lawsuits, then your decisions about which terms to extract will differ from those needed by someone writing a technical manual on centrifuge parts and use. Do you need product names or just specialized processes? Are you going to include campaign slogans or boilerplate text for the sake of practicality even if they are not true terms?

Extract terms that align with aims/goals	If you are trying to standardize your forestry terminology, for example, then you could omit extracting terms from other fields you come across in the corpus. If you came across specialized terms relating to butterfly breeding in the forestry corpus, you could either skip extracting them or remove them from your term candidate list after extraction. On the other hand, if you are going to translate texts relating to forestry laws, law terms related to forestry should be extracted too.
Domain specificity vs general lexicon	Will you focus efforts on only technical terms, or will you include some general language words you could find in any dictionary?
Visibility of the term/legal issues	Do you need to protect your brand? Brands, slogans, and trademarked terms all need to be used correctly to establish and maintain ownership. To ensure uniform use, document these. For example, if your brand name always needs to be presented in italics, specify this in the terminology record in a usage note.
Translation difficulty/obscure terms	Did it take a long time to find or decide on a term? Consider cataloging terms and related data that took a good deal of resources (time and effort) to research. Just like bookmarking a web page so you don't have to search for it again, you can avoid repeatedly researching terminology options or debating which term is to be used and which is not the next time this case comes up. Research it once, document the results/decisions in the terminology database and make that data available to all. Consider it a duty to the common good, or at least a help to your future self.
Homographs and polysemous terms (same form of word with multiple possible meanings)	Will you include homographs and ensure you avoid potential semantic ambiguity and confusion around terms that look the same or terms with multiple meanings? For polysemous terms, the best practice is to respect the single-concept principle and create one record per concept, rather than one record with multiple meanings for the same term form. You can still offer a dictionary-style view for users if you like.
Acronyms	Have you ever come across an acronym and had no idea what it meant because the author did not take the time to provide its decoded form, but you did not know whom to ask for clarification? Knowing the correct long form and correct short form of a term is important, particularly with in-house acronyms. For new employees, having a central repository of these in-house terms along with what they mean can help them get up to speed faster than they would otherwise without having to pester their supervisor. To find them in a corpus, sometimes it helps to search for parentheses.

Synonyms	Will you make one synonym the preferred or obligatory term over another? Often, you can find them by searching for language patterns such as “also known as,” “also called,” and “another word for.”
Potentially controversial terms	How much risk can you handle? Consider recording geopolitically sensitive terms (for example, country names) and noninclusive terms (biased by gender, race, sexual orientation, etc.) to be avoided, and which terms to use instead. This is useful for avoiding problems when localizing/selling in some countries, or for preventing offense to existing or potential customers.
Forbidden terms	Are there any terms you may not use under any circumstances? If yes, document them in the terminology database.
New terms (neologisms) or new product names	Do you want to target new terms? Searching out and documenting new terms can help you flag terms that may be difficult to translate or where the choice of translation has not yet been settled. Recording them can help spread the use of those new terms in your organization.
Any term that needs to be used in the same way by everyone	Do you use controlled language? Do you have similarly named official titles for positions or administrative divisions? Do you ever need to know the full name of a law, not just its informally used name? If yes, then you might decide to include these kinds of terms in your terminology database.
Finding boundaries of multiword terms	How will you know where a term ends, particularly for multi-word terms? For example, if you see terms in a sentence and need to determine what to extract, you may have some debates as to where the term ends. It may not always be clear. For example, in French (**réglage pneumatique de pression de jointage)** or English (**pneumatic adjustment of jointing pressure**), or even in German where the boundaries are often more easily identified (**Pneumatische Fugendruck-Verstellung)**, this may be challenging. To help you determine where a term ends, ask yourself, if I stop here, is this term still describing a unified concept? You may also find more than one term in the same segment of text. For example, in **acoustic warfare signal processing system** you find at least two other terms: **acoustic warfare** and **signal processing.**
Frequently occurring terms	How will you know whether a term is important to extract? A term occurring often in a text is a big hint that it’s important and worth extracting. However, sometimes critical terms only occur once, so terminology extraction tools that only extract terms that appear at least twice will not extract them. For example, to extract terms from software interfaces, you might need an extraction method that can identify terms even if only present once (with a tool or manual extraction by a person).

Learn more: See the Pavel Terminology Tutorial 2.5.2 Recognition of Terminological Units for information and practice.

Where should I look for terms?

You need to submit a suitable "corpus" to the TE tool. A corpus is a selection of documents in electronic file format. The format needs to be readable by the TE tool. Check the documentation of your TE tool to determine the supported file formats.

Binary files
Binary files are not parsable. Tools that say they handle documents in various common formats such as PDF and Office formats (Excel, PowerPoint, MSWord, and so forth) usually are converting these under the hood.
Examples: PDF, jpg, optically scanned documents

Nonbinary files
This usually includes file types such as .txt, .html, or .xml.
A quick test to determine whether a file is binary (and therefore unsuitable) or not is to open the file in a text editor such as Notepad. If the contents are readable in Notepad, then the file is not binary.

Typically, you need to assemble the corpus by collecting the files and organizing them in one location such as a folder (directory) on a computer (or in the cloud), from which the files can be uploaded to the TE tool. Some tools allow the submission of only one file at a time; if this is the case and if your corpus contains multiple files (as they typically do), all files will need to be merged into one file. There are utilities available on the internet that can do this merging task automatically. Some tools allow you to upload files from a directory. Some support the submission of a .zip file. And finally, some tools require you to submit a "list" of the files that includes the name and full path (location) of each file.

The corpus should contain as many files as possible; results from a large corpus are typically better than results from a small corpus.

Select files for the corpus that reflect the purpose of the extraction. For example, if you are interested in a certain subject field, select only files that contain material that is related to that subject field. If you are interested in terms that are used in a product's user interface, select the files that contain those UI strings, as well as the related documentation and online help files. If you need to extract terms relating to a new company product, collect all the files describing the product. If you are doing a bilingual extraction, you will need bilingual files such as XLIFF or TMX, or pairs of unilingual files. On the other hand, if you are doing a monolingual extraction and the only file format you have is bilingual (XLIFF or TMX), you may have to remove the target language strings. You can do this with an advanced text editor that supports regular expressions, or an XML editor that supports XPath expressions.

Removal of target language strings is one example of a preprocessing task that may be required prior to running the TE tool. You should always examine the corpus files before running the tool to verify that there is no corruption (particularly for accented languages), to ensure that they are in the proper encoding format (usually UTF-8 is recommended, but this depends on the tool and the language), and to identify possible markup that the tool may have difficulty with. Often it makes sense to run a trial extraction first and look at the output. Problems are usually quite evident; they can then be addressed in the source files and the extraction can be run again. The tool may have difficulty, for example, with certain HTML or XML markup, which can subsequently be removed and the tool run again.

Keep copyright issues in mind if you intend to use other people's documents, glossaries, databases or other resources to extract context sentences for your database (in contrast with just extracting terms). You may need to obtain permission first, for example if you intend to repurpose a large percentage of the content for commercial purposes. You may also need to cite your sources in your terminology records.

Where can I look for terms?

- In bitexts
- In translation memory files (TMX)
- In my organization's documentation
- In terminology-rich documents online or in databases (glossaries, articles, manuals, technical specifications, monographs, patents, standards, technical documents, laws, dictionaries, encyclopedia)
- In paper documents (either highlight and transcribe terms manually or have documents digitized to be run through term extraction software)
- In software UI string export files
- In existing corpora

Walk-through of term extraction steps

Note that the process may vary depending on your tools and on your goals.

1. Define why term extraction is needed, as well as how and by whom terms will be used (communication, marketing, legal, localization, translation, etc.).
2. Determine the parameters of the term extraction (deadline, type of term extraction, subject field, languages, number of terms desired and type of terminology project). Examples: Manually selecting 100-word glossary of terms and contexts in Swahili on a particular narrow subject field; or automated extraction of terms to be validated by a terminologist or terminology worker for feeding a 25,000-term terminology database in multiple languages and covering multiple subject fields for organization-wide access.

3. Set criteria for selecting documents to extract from, and select documents to form part of the extraction corpus accordingly.
4. If the documents are not already in useful formats, convert the files into the formats your term extraction tool can handle. If aligning pairs of documents with an alignment tool for bilingual extraction, convert the files to the formats it can handle.
5. If planning to perform bilingual extraction, use either a dedicated alignment tool or the one integrated in your term extraction tool to align pairs of documents into a bitext or TMX.
6. Set the parameters for your term extraction if the tool allows (minimum frequency of terms to be extracted, minimum and maximum length of terms, etc.).
7. Load the files and run the term extraction process.
8. Examine the resulting list of term candidates – if there are too many candidates to examine, you could either focus on the higher-frequency results or repeat step 6 with a higher minimum-frequency setting and rerun the extraction. In a case where you are planning a 300-word glossary and the tool generates only 40 candidates, repeat step 6 with a lower minimum-frequency setting.
9. Validate the list of term candidates to add the desired terms (and related information such as contexts, if possible) to a terminology database, and discard out-of-scope, irrelevant or non-terms. This step may also be useful for identifying typos in your corpus (see the lowest-frequency term candidates).
10. If desired and if your tools allow it, batch query that list of term candidates against your existing database content to ascertain what you already have and what you don't. This step will help you determine whether to upload the entire list or not, as well as find gaps in your terminology base coverage and identify which terms you need to add.
11. Flesh out and validate the resulting terminology records.

How to clean the raw output of term extraction

- Check for termhood – Is the term candidate really a term according to your criteria?
- Check term boundaries – Where does my term start and end?
- Keep term families in mind – Are the concepts that my terms represent related?
- Filter out or flag terms already in the termbase – Do I already have records for these terms in my terminology database?
- Reduce terms to their canonical form – Change plurals to singulars, conjugated verbs back to the base form, etc.
- Change capitalization for proper nouns if the extractor puts all terms in lower case; change to lower case if extracting terms from capitalized headings and titles.
- Stay within subject field boundaries – If creating a glossary for a given subject field, skip terms not specific to that field.
- Watch for frequently reused term components – Is there a shorter term within a longer term candidate? That could signal a group of related terms. For example, if you see *thermosphere, stratosphere, exosphere,* and *mesosphere* with a repeating element (-sphere), that's a clue that they are related terms. Also, if you see term candidates like *automatic gearbox, manual gearbox, gearbox casing, gearbox output shaft, dual-clutch gearbox, and planetary gearbox*, you may safely assume that

gearbox should also be extracted as a term, even if it does not show up by itself in your term candidates list.

Elements to consider when selecting terms from list of term candidates	**Candidates to exclude**	**Candidates to include**
Termhood	● who	● World Health Organization ● WHO
Term boundary	● leave without ● leave without pay for	● leave without pay ● leave without pay for care of immediate family
Canonical form	● bed bolts ● intubated ● out-of school children ● vulnerable people ● vulnerable persons	● bed bolt ● intubate ● out-of-school child ● vulnerable person ● Office of the Vulnerable Persons' Commissioner
Capitalization and numbering	● 5.3 Executive Committee Meeting	● executive committee meeting
Subject field boundaries (for this example, if selecting terms in the field of virtual reality)	● 6-inch howitzer shell ● spotfin butterflyfish ● main transmission gearbox	● augmented reality ● extended reality ● social virtual reality ● spatial augmented reality
Term families (if working with concept networks – trees, for example, in this case)	● marigold ● beavertail ● mosquito	● maple ● oak ● elm ● fir ● sequoia
Terms within longer term candidates	● upcoming abrogation of the Personal Information Protection and Electronic Documents Act	● personal information ● electronic document ● Personal Information Protection and Electronic Documents Act ● abrogation

How noise and silence affect term extraction

No tool is perfect – there will be some noise or some silence, or some of both.

Noise = The items in the results that are not desired or useful (extra items included that I don't want)

For example, an extractor that offers me the word “and” as a potential term is offering me noise.

Silence = The items not included in the results that would have been useful had I seen them (missed items)
For example, if the text contains the acronym, “A.N.D.” but the extractor does not include it in the list of terms for me to validate, that would be silence.

Depending on the tool, and sometimes depending on the settings, noise and silence vary. Generally, a good tool minimizes both. Ideally, the tool would have a default setting somewhere in the middle, but it would allow users to modify the settings depending on the project’s needs. For example, if I need to pull the 100 terms used most often in a subject field from a shorter document, then I might lower the minimum number of times a term has to appear in the document before the extractor presents it to me. Conversely, if I have a large corpus, I might raise that threshold in order not to have to sift through term candidates that appear only rarely and are unlikely to be essential concepts.

It’s about managing risk (If there are too many silences, which important terms will I not see?) and managing time (With too much noise, will it take too long to go through my list of candidate terms to find the true, useful and pertinent terms?). Which would you rather do, look at a list of 100 terms and keep 90, perhaps having missed another 50 important terms, or spend more time looking at a list of 1,000, keeping 300, yet feeling more confident you missed only a couple of important terms? Your choice may change depending on the goal of your project and the length of your deadline.

Noise = unwanted results
Silence = missed terms
Trade-off unavoidable

Using a concordance tool to do more research (in addition to term extraction)

Concordance functionality may be offered by some term extraction tools. If not, you may wish to use a separate concordance tool. Concordance searches of your term extraction corpus files can help you find:

- salient one-word terms (unigrams),
- complex multiword terms (for example, “Personal Information Protection and Electronic Documents Act”), and
- collocates (verbs or adjectives that are commonly or exclusively used with a term – for example, abrogate a law, enact legislation).

KWIC concordance helps identify collocations or technical verbs and related terms – KWIC means Key Word in Context. For example, a search of a text for the term **polymerization** might return something looking like this:

… polymer or resin made by	**polymerization**	of chemical compounds containing the...
…reaction usually occurring during chain	**polymerization**	, in which an active macromolecule…
…is useful to note: chain	**polymerization**	is desirable in this case.

From examining the words preceding the term on the left, we can see that **chain polymerization** is worth examining as another term we could extract. Also examining the words to the right of term can point to other terms to extract.

When selecting a TE tool, what should you be wary of, and why?

Characteristic	**Reason**
TE tools that support all languages	They are statistical and often don't provide the quality of one tailored to a language's common term patterns.
TE tools that lock you into an over-arching software solution (CAT, CA)	Such tools make it hard to share data and change tools later.
Overestimating the ability of bilingual extraction to find the right equivalent	Target terms all have to be manually checked.
TE tools that force the output into a single file format, such as MSWord	You may have trouble sharing data with others if their system does not import that file format.
TE tools that lock you into a preset extraction, not allowing you to change the parameters of your extractions.	It is important for the tool to offer flexible and customizable extraction options so that you can tailor your extraction to the text and your needs. Better tools let you modify the minimum occurrence levels, modify stop-word lists, limit term length and more.
TE tools that do not let you access the term candidates in context.	Being able to see the candidate term in context helps you determine whether it is a real term and how it is used.
TE tools that do not let you select terms from the texts that you come across just because they were not in the list	You want the flexibility to manually add terms you stumble across in contexts or in other terms, because no tool is perfect.
TE tools that don't export data in properly structured formats.	This type of problem may cause import issues with some termbases, requiring either pre-import conversion or post-import cleanup work to ensure data elements are separated before you can begin.

TE tools that won't allow users to export just the term candidates list	If you work with several TE tools, selecting the best one for each project, you may wish to export just the term candidates list for further work by someone else or in another tool. For example, an entry-level worker may have enough skill to assemble a corpus and run it through the TE tool, and could then pass the list of candidate terms to a language professional to judge which candidate terms should be retained before any are entered into the termbase.
TE tools that can handle only one file type	You may have to spend time converting the files to the allowed input format before you can begin.
If you have security concerns or sensitive texts, TE tools that reuse or resell your corpus, making you cede all rights to texts in perpetuity	Your organization may have legal obligations to protect private information or may want to protect proprietary information. Always check the Terms and Conditions as well as the End-User-License-Agreement carefully. Free tools may have hidden costs.
If you have security concerns or sensitive texts, TE tools that store your documents on a server in a country that has laws allowing that country's government access to those files	You may wish to ensure that no one outside of your company has a legal right to access documents that are in your corpora. Find out where the data is stored, and check the privacy and other applicable laws.

Term extraction scenarios: Advice from TerminOrgs members

Sometimes, using term extraction is vital. Other times, using terminology extraction tools is not the appropriate solution. Consider these scenarios.

Scenario 1: A department has no terminology at all.
Advice: Collect 50 documents from the recent past and do a term extraction.
Sort the results by term saliency to find the most specific terms for your subject field.

Scenario 2: A department has Excel lists but is not sure whether the terms in them match with their existing texts
Advice: Do a term extraction of 20 recent documents and compare/match the results with your existing Excel list.

Scenario 3: A department has a termbase, but the quality is poor.
Advice: Before considering term extraction, spend the time to improve your existing terminology records (add acronyms, definitions and pictures; group synonyms to the concepts; and add a term status). From there, move on to term extraction if needed.
Advice: You can use the corpus search or KWIC functions in your TE tool or translation memory tool to search the documents for patterns that can indicate synonyms (for example, *aka, called, known as, another name for, another word for*), abbreviation forms (for example, *short for* or an acronym) or definitions (for example, *is a, is defined as, described as*).

Scenario 4: A department does translations into other languages.
Advice: Extract terms from the source text to be translated and match the results against your existing termbase. Integrate term extraction of the source text into your regular translation workflow.

Scenario 5: A department has only a few short paper documents to scan for terms for tomorrow.
Advice: Either take an hour to digitize them and then run them through the extractor, or use a highlighter to scan manually and transcribe the data manually.

Scenario 6: A department has older electronic files in which the terms are embedded in images or diagrams.
Advice: Scan the documents manually.

Scenario 7: A department has documents in a format the term extraction tool can't handle.
Advice: Either batch convert the electronic files to a format it can handle; extract terms manually (if the volume is not too large and your deadline permits); determine whether the documents are core documents (if no, set them aside and focus on core documents); or buy a new extraction tool that can handle those formats.

Scenario 8: A department has a large set of documents to translate that are updated each year, with new terms popping up every year.
Advice: Create a bilingual corpus of last year's documents and their translations. Align them. Have your term extractor extract the terms and equivalents and load them into the termbase, always with the validation of a language professional. When the new version of the source text arrives, run unilingual extraction on it, and then batch search that list of candidate terms against your termbase. Where an important candidate term is not found in the termbase, focus your terminology research for equivalents of them. (Also, add the aligned corpus to your translation memory tool, if you have one, so that unchanged segments from previous years can be reused.)

Scenario 9: A large set of documents on the same topic needs to be split among several translators because it is too large for one translator to finish before the deadline, and the various slices will be cobbled back together before delivery.
Advice: Create a corpus from the set of documents to be translated. Obtain the list of term candidates with the extractor, and batch search your termbase for those records. Ensure that all the terminology records found are made centrally available to all translators working on that project. Invest some research up front into researching those not found so that records can be quickly made and shared with the translators. This way they can use consistent terminology right away in all slices of their translations, instead of having to go back and fix terminology inconsistencies later, either just before delivery, or worse, after a client complaint.

Other resources on term extraction or extractors

http://www.atanet.org/chronicle-online/wp-content/uploads/4407_19_inge_karsch.pdf
https://termcoord.eu/free-term-extractors/
http://bikterminology.com/?s=term+selection