

# New Client, New Terms: Glossary 101

---

Session 066

ATA 61st ANNUAL CONFERENCE  
Friday, Oct. 23, 2020, 5:00 p.m. EDT



ARGOSMULTILINGUAL.COM

Karl Pfeiffer, CT  
**Senior Language Lead**  
karl.pfeiffer@argosmultilingual.com

# ABOUT THE PRESENTER



**Karl Pfeiffer, CT**  
ATA-certified  
English>German  
translator

After earning a degree in physics from the University of Tübingen, Germany, Karl began translating engineering documents. He is now a Senior Language Lead at Argos Multilingual, a global language solutions provider with offices in Krakow, Poland, and Boulder, Colorado.

He focuses on the management of linguistic assets (glossaries, translation memories, etc.), workflow optimization, and quality enhancements.

With decades of experience as a freelancer as well as an in-house professional, Karl brings the perspectives of the bilingual translator and the multilingual needs of a leading language service provider to the discussion.



# THE CHALLENGE

- ✔ You've landed a contract with a new direct or corporate client.
- ✔ You are adding a new subject matter domain or product line to an existing portfolio.
- ✔ There is no prior glossary, or no approved glossary.
- ✔ Knowing that the initial setup of a glossary will directly impact downstream processes, from translation and revision to the quality assurance (QA) phase and in-country review, **what is the best strategy for establishing a new glossary?**



# THE BENEFITS

- ✓ **Authoring stage:** supporting the goal of unambiguous, clear source text
- ✓ **Translation and revision workflow:** consistent rendering of terms in the target language
- ✓ **QA workflow:** efficient deployment of QA tools, reducing false or missed positives
- ✓ Terminology work may be one of the last areas of the translation process where **human involvement** still matters more than automation but neglecting it may be costly.

# THE BASICS



**Characteristics of a  
glossary term**



**Nomenclature**

- ✓ Concept
- ✓ Term
- ✓ Definition
- ✓ Other metadata



# THE BASICS



## Characteristics of a glossary term

- ✓ A **term** is associated with a specialized concept in a particular subject field.
- ✓ Terminology is the study of the **concepts** and terms belonging to specialized languages.
- ✓ A **term** may be composed of one word or two or more words.



# THE BASICS



## Nomenclature: Research Principles

- ✓ **Concepts** group individual objects into classes by means of shared characteristics or features.
- ✓ Concepts are units of knowledge identifiable by their stable association with a set of **semantic features** and with one or more **designations (terms)**.
- ✓ Essential semantic features help position concepts in a **conceptual system**.
- ✓ **Stable links between concepts, semantic features and terms** help identify and define concepts based on textual matches present in the specialized documentation.



# THE BASICS



## Nomenclature: Terminology records

- ✓ **Term extraction:**
  - identify candidate terms that are pertinent to the field of research
  - find textual supports that provide information about the related concepts
- ✓ **Textual supports:** information about the concept and about term usage
- ✓ **Definitions:** include essential and delimiting semantic features of the concept being defined
- ✓ **Contexts, usage samples and phraseologisms** (fixed expressions, such as idioms and other types of multi-word lexical units) show how terms are actually used in a specialized language





# THE BASICS



## Nomenclature: Terminology records

### Other metadata:

- ✓ *lexical basics*: part of speech, gender, number
- ✓ *usage labels*: usage particularities associated with the terms
- ✓ subject field or subfield
- ✓ source references
- ✓ originator and updater names
- ✓ creation and update dates
- ✓ approval status



## THE STEPS

**E**xtract terms

**L**abel with metadata

**A**nalyze and review for consistency, suitability, and scope

**T**ermbase import and maintenance

**D**istribute to source and target teams



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Manual analysis by selective parsing of the source documents



Batch analysis by integrated and third-party extraction tools



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY

Two sample source documents:

---

**DIRECTIVE 2011/83/EU OF THE EUROPEAN PARLIAMENT  
AND OF THE COUNCIL**

**of 25 October 2011**

**on consumer rights, amending Council Directive 93/13/EEC and  
Directive 1999/44/EC of the European Parliament and of the  
Council and repealing Council Directive 85/577/EEC and Directive  
97/7/EC of the European Parliament and of the Council**

(<https://eur-lex.europa.eu/eli/dir/2011/83/oj>)



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY

Two sample source documents:

---

**REGULATION (EU) 2016/679 OF THE EUROPEAN  
PARLIAMENT AND OF THE COUNCIL  
of 27 April 2016  
on the protection of natural persons with regard to the processing  
of personal data and on the free movement of such data, and  
repealing Directive 95/46/EC (General Data Protection Regulation)**

(<https://eur-lex.europa.eu/eli/reg/2016/679/oj>)



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Manual analysis by selective parsing of the source documents

a) Parsing the Table of Contents

☰ Hide Table of contents

[Top](#)

[CHAPTER I - General provisions](#)

[CHAPTER II - Principles](#)

[CHAPTER III - Rights of the data subject](#)

[Section 1 - Transparency and modalities](#)

[Section 2 - Information and access to personal data](#)

[Section 3 - Rectification and erasure](#)

[Section 4 - Right to object and automated individual decision-making](#)

[Section 5 - Restrictions](#)

[CHAPTER IV - Controller and processor](#)



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Manual analysis by selective parsing of the source documents

a) Parsing the Table of Contents

 Hide Table of contents

Top

CHAPTER I - SUBJECT MATTER,  
DEFINITIONS AND SCOPE

CHAPTER II - CONSUMER INFORMATION  
FOR CONTRACTS OTHER THAN DISTANCE  
OR OFF-PREMISES CONTRACTS

CHAPTER III - CONSUMER INFORMATION  
AND RIGHT OF WITHDRAWAL FOR  
DISTANCE AND OFF-PREMISES CONTRACTS

CHAPTER IV - OTHER CONSUMER RIGHTS

CHAPTER V - GENERAL PROVISIONS

CHAPTER VI - FINAL PROVISIONS

ANNEX I

ANNEX II



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Manual analysis by selective parsing of the source documents

b) Parsing Definitions

<a href="#">Hide Table of contents</a>
<a href="#">Top</a>
<a href="#">CHAPTER I - SUBJECT MATTER, DEFINITIONS AND SCOPE</a>
<a href="#">CHAPTER II - CONSUMER INFORMATION FOR CONTRACTS OTHER THAN DISTANCE OR OFF-PREMISES CONTRACTS</a>
<a href="#">CHAPTER III - CONSUMER INFORMATION AND RIGHT OF WITHDRAWAL FOR DISTANCE AND OFF-PREMISES CONTRACTS</a>

*Article 2*

## **Definitions**

For the purpose of this Directive, the following **definitions** shall apply:

- (1) **'consumer'** means any natural person who, in contracts covered by this Directive, is acting for purposes which are outside his trade, business, craft or profession;
- (2) **'trader'** means any natural person or any legal person, irrespective of whether privately or publicly owned, who is acting, including through any other person acting in his name or on his behalf, for purposes relating to his trade, business, craft or profession in relation to contracts covered by this Directive;
- (3) **'goods'** means any tangible movable items, with the exception of items sold by way of execution or otherwise by authority of law; water, gas and electricity shall be considered as goods within the meaning of this Directive where they are put up for sale in a limited volume or a set quantity;





# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Manual analysis by selective parsing of the source documents

## b) Parsing Definitions

☰ Hide Table of contents

Top

CHAPTER I - General provisions

- Article 1 - Subject-matter and objectives
- Article 2 - Material scope
- Article 3 - Territorial scope
- Article 4 - Definitions

CHAPTER II - Principles

CHAPTER III - Rights of the data subject

### Article 4

#### Definitions

For the purposes of this Regulation:

- (1) 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;
- (2) 'processing' means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction;
- (3) 'restriction of processing' means the marking of stored personal data with the aim of limiting their processing in the future;



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Manual analysis by selective parsing of the source documents

- c) Parsing the included index, appendix, and glossary  
(if available)
- d) Parsing list of tables and figures  
(if available)



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Batch analysis by integrated and third-party extraction tools



SDL PHRASEFINDER



Knowledge Base

## How to extract terminology



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



## Batch analysis by integrated extraction tools (SDL PhraseFinder)



SDL PHRASEFINDER

### Legacy tools:

*SDL MultiTerm Extract*

- from Trados side
- mathematical functionality ("if word A always appears in sentences for which word B always appears in the translated sentence, then these words must form a word pair")

*SDL PhraseFinder*

- SDLX companion
- works on a language-based level for English, French, German, Spanish, Dutch, and Portuguese
- less accurate for other languages



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Batch analysis by integrated extraction tools (SDL PhraseFinder)

	A	B	D	E	G	H	I
1	SOURCE [ORIGINAL]	SOURCE [ROOT]	CONTEXT	CONTEXT-FILENAME	FREQUENCY	RANKING	PART OF SPEECH
3	archiving purposes	ARCHIVING PURPOSE	Further processing for archiving purposes in the public inte	GDPR_EN.docx_en-US	19	8	Noun Phrase
5	public health	PUBLIC HEALTH	It should also be for Union or Member State law to determin	GDPR_EN.docx_en-US	14	8	Noun Phrase
6	Union law	UNION LAW	National authorities in the Member States are being called u	GDPR_EN.docx_en-US	14	8	Noun Phrase
7	public security	PUBLIC SECURITY	The protection of natural persons with regard to the processir	GDPR_EN.docx_en-US	13	8	Noun Phrase
12	contact details	CONTACT DETAIL	This Regulation does not cover the processing of personal	GDPR_EN.docx_en-US	12	8	Noun Phrase
13	criminal convictions	CRIMINAL CONVICTION	The risk to the rights and freedoms of natural persons, of	GDPR_EN.docx_en-US	12	8	Noun Phrase
14	Data Protection Supervisor	DATA PROTECTION SUPERVISOR	... and the European Data Protection Supervisor or ...	GDPR_EN.docx_en-US	12	8	Capitalization; Noun Phrase
15	Protection Supervisor	PROTECTION SUPERVISOR	It should consist of the head of a supervisory authority of eac	GDPR_EN.docx_en-US	12	8	Capitalization; Noun Phrase
18	issue guidelines	ISSUE GUIDELINE	The Board may also issue guidelines on processing operat	GDPR_EN.docx_en-US	11	8	Noun Phrase
19	professional secrecy	PROFESSIONAL SECRECY	by persons subject to a legal obligation of professional se	GDPR_EN.docx_en-US	11	8	Noun Phrase
20	criminal penalties	CRIMINAL PENALTY	The protection of natural persons with regard to the processir	GDPR_EN.docx_en-US	10	8	Noun Phrase
21	medium-sized	MEDIUM-SIZED	In order to ensure a consistent level of protection for nat	GDPR_EN.docx_en-US	9	8	Adjective; Noun Phrase
22	Official Journal	OFFICIAL JOURNAL	Official Journal of the European Union; Official Journal of the	GDPR_EN.docx_en-US	8	8	Capitalization; Noun Phrase

941	war crimes	WAR CRIME	Member States should also be authorised to provide fo		1	8 646	war crimes
942	well-being	WELL-BEING	This Regulation is intended to contribute to the accom		1	8 647	well-being
943	widespread public	WIDESPREAD PUBLIC	The objectives and principles of Directive 95/46/EC ren		1	5 648	widespread public




# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Batch analysis by integrated extraction tools

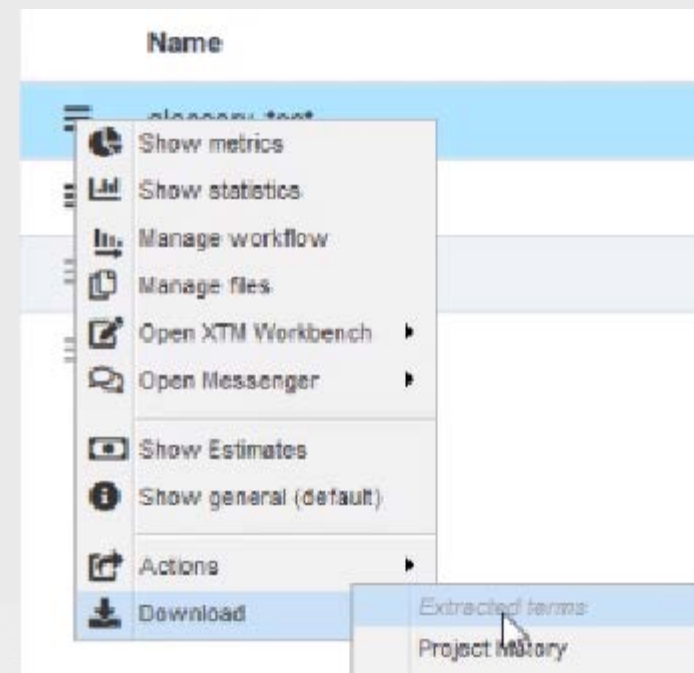


1. Enable terminology extraction by clicking on the Configuration tab  > Settings > Translation > Terminology, and checking the box in the Permit column next to

Run Terminology extraction



2. After the project is created and file analysis is complete, go to the Projects tab > Projects and select Download > Extracted terms from the dropdown menu of the project.



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Batch analysis by integrated extraction tools



	A	B	C	D	E	F
1	en_US	Surface Forms	Frequency	Ignore	Found in file	Context
	author	authority (951) authorities (180) author (1)	1132		GDPR_EN.doc	Notwithstanding paragraph 1, Member State law may require controllers to consult with, and obtain p
2	process	processing (1062) process (24) Processing (20) processes (4)	1110		GDPR_EN.doc	Where the controller <b>processes</b> a large quantity of information concerning the data subject, the contr
3	control	controller (952) controllers (60) control (15) Controllers (2) Controller (1)	1030		GDPR_EN.doc	In order to ensure a consistent level of protection for natural persons throughout the Union and to pre
4	member	Member (812) members (37) member (25)	874		GDPR_EN.doc	It should be possible to entrust supervision of such data processing operations to specific bodies wit
5	law	law (442) lawful (25) laws (11) lawfulness (9) Lawfulness (1)	488		GDPR_EN.doc	To that extent, this Regulation does not exclude Member State <b>law</b> that sets out the circumstances
6						



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Batch analysis by integrated extraction tools



A	B	C	D	E	F
process	processing (1062) process (24) Processing (20) processes (4)	1110		GDPR_EN.doc	<p>Where the controller <b>processes</b> a large quantity of information concerning the data subject, the controller should be able to request that, before the information is delivered, the data subject specify the information or <b>processing</b> activities to which the request relates.</p> <p>-----</p> <p>'<b>processor</b>' means a natural or legal person, public authority, agency or other body which <b>processes</b> personal data on behalf of the controller;</p> <p>-----</p> <p>If the purposes for which a controller <b>processes</b> personal data do not or do no longer require the identification of a data subject by the controller, the controller shall not be obliged to maintain, acquire or <b>process</b> additional information in order to identify the data subject for the sole purpose of complying with this Regulation.</p> <p>-----</p> <p><b>processes</b> the personal data only on documented instructions from the controller, including with regard to transfers of personal data to a third country or an international organisation, unless required to do so by Union or Member State law to which the <b>processor</b> is subject; in such a case, the <b>processor</b> shall inform the controller of that legal requirement before <b>processing</b>, unless that law prohibits such information on important grounds of public interest;</p>
Processor	processor (217) processors (46) Processor (1)	264		GDPR_EN.docx	Any processing of personal data in the context of the activities of an establishment of a controller or a <b>processor</b> in the Union should be carried out in accordance with this Regulation, regardless of whether the processing itself takes place within the Union.





# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Batch analysis by integrated extraction tools



495	urgent opinion	urgent opinion (3)	3	GDPR_EN.docx
496	user	users (2) user (1)	3	GDPR_EN.docx
497	VIII	VIII (3)	3	GDPR_EN.docx



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



## Batch analysis by third-party extraction tools (Sketch Engine)

---

- developed by Lexical Computing Limited since 2003
- cloud-based corpus manager and text analysis software
- corpora in 90+ languages
- named after one of the key features, **word sketches**: one-page, automatic, corpus-derived summaries of a word's grammatical and collocational behavior



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Batch analysis by third-party extraction tools (Sketch Engine)

The screenshot displays the Sketch Engine interface for the word "team" in the English Web 2015 (enTenTen15) corpus. The main view shows four panels of related terms, each with a table of results. The panels are: "nouns modified by 'team'", "modifiers of 'team'", "verbs with 'team' as subject", and "verbs with 'team' as object". Each panel includes a table with columns for the term, frequency, and a score. The interface also features a search bar, a sidebar with navigation icons, and a top navigation bar with various tool icons.

nouns modified by "team"			
member	229,738	9.71	...
team members			
leader	65,274	8.79	...
team leader			
captain	12,413	8.02	...
team captain			
player	26,140	7.99	...
a team player			
sport	13,842	7.82	...
team sports			
mate	10,187	7.76	...
team mates			
spirit	13,350	7.7	...

modifiers of "team"			
management	88,815	8.35	...
management team			
football	62,912	8.31	...
football team			
project	77,699	8.24	...
project team			
research	96,161	8.12	...
research team			
leadership	58,236	8.11	...
leadership team			
basketball	46,175	7.94	...
basketball team			
development	60,570	7.72	...

verbs with "team" as subject			
win	37,701	8.49	...
team won			
work	65,490	8.37	...
develop	28,761	7.96	...
play	28,612	7.87	...
consist	21,993	7.8	...
team consists of			
compete	14,106	7.44	...
teams competing			
have	305,225	7.37	...
team has			
take	31,778	7.14	...

verbs with "team" as object			
join	119,809	9.83	...
lead	106,791	9.41	...
contact	22,909	8.11	...
form	26,817	8.05	...
manage	22,169	7.66	...
assemble	11,499	7.59	...
coach	9,596	7.41	...
head	9,650	7.26	...
team headed by			
build	27,884	7.23	...



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



a word	a phrase	text
<p>✓ <b>TYPICAL COMBINATIONS</b> type a word and see a <a href="#">Word Sketch</a> with the most typical combinations, i.e. <a href="#">collocations</a> of the word</p>	<p>✓ <b>TYPICAL COMBINATIONS</b> type a multi-word phrase, such as <i>good idea</i> and see a <a href="#">Word Sketch</a> with other words that typically appear together with the phrase</p>	<p>✓ <b>GENERATE A</b> generate a <a href="#">word list</a> of the most frequent or even all words, nouns, adjectives, words beginning/ending with... etc. Set your own criteria and output options.</p>
<p>✓ <b>SYNONYMS</b> type a word and see a <a href="#">thesaurus</a> with synonyms and similar words</p>	<p>✓ <b>FIND EXAMPLES IN CONTEXT</b> look up examples of the phrase <a href="#">in context</a> or search for <a href="#">patterns</a> without specifying concrete words</p>	<p>✓ <b>EXTRACT KEY WORDS AND</b> use Sketch Engine to extract <a href="#">key words or terminology</a>, that is typical for your text or web</p>
<p>✓ <b>COMPARE WITH ANOTHER WORD</b> type two similar words and observe the differences in use by <a href="#">comparing their collocations</a></p>	<p>✓ <b>LOOK UP TRANSLATIONS</b> use <a href="#">parallel concordance</a> to look up examples of how others translated the phrase</p>	<p>✓ <b>BILINGUAL TERMINOLOGY</b> <a href="#">extract terms</a> and their translation from parallel texts to create a glossary of terminology</p>
<p>✓ <b>FIND EXAMPLES OF USE</b> look up examples of the word <a href="#">in context</a> as used by real users of the language and even filter <a href="#">easy-to-understand examples</a></p>		<p>✓ <b>CALCULATE</b> generate a list of the most frequent (or all) <a href="#">multi-word expressions</a> (or MWUs) such as <i>at the beginning of, in the form of or in relation to the</i></p>
<p>✓ <b>LOOK UP TRANSLATION</b> use <a href="#">parallel concordance</a> to look up examples of how others translated the word.</p>		<p>✓ <b>IDENTIFY NEOLOGISMS</b> perform <a href="#">diachronic analysis</a> to discover neologisms or words going out of use</p>
		<p>✓ <b>TAG TEXT FOR PARTS OF SPEECH</b> have all the words in your text tagged for part of speech (<a href="#">POS tagging</a>), this happens automatically each time you <a href="#">upload text</a></p>



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Batch analysis by third-party extraction tools (Sketch Engine)

**DASHBOARD**

**GDPR\_EN**

- Word Sketch**  
Collocations and word combinations
- Word Sketch Difference**  
Compare collocations of two words
- Thesaurus**  
Synonyms and similar words
- Concordance**  
Examples of use in context
- Parallel Concordance**  
Translation search
- Wordlist**  
Frequency list
- N-grams**  
Multiword expressions (MWEs)
- Keywords**  
Terminology extraction
- Trends**  
Diachronic analysis, neologisms
- Text type analysis**  
Statistics of the whole corpus
- OneClick Dictionary**  
Automatic dictionary drafting



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Batch analysis by third-party extraction tools (Sketch Engine)

The screenshot shows the 'KEYWORDS' section of the Sketch Engine interface. At the top, the word 'KEYWORDS' is displayed in large blue letters. To its right is a search input field containing 'GDPR\_en', a search icon, and an information icon. Below this are three tabs: 'BASIC' (selected), 'ADVANCED', and 'ABOUT'. The main content area has a light blue background and contains the following text:

Keywords and terms help us understand what the topic of the corpus is or how it differs from the reference corpus. By default, general language corpora are used as reference corpora to represent non-specialized language.

<b>Keywords</b> individual words (tokens) which appear more frequently in the focus corpus than in the reference corpus.	<b>Terms</b> multi-word expressions which appear more frequently in the focus corpus than in the reference corpus and, additionally, match the typical format of terminology in the language.
---	--



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



## Batch analysis by third-party extraction tools (Sketch Engine)

BASIC **ADVANCED** ABOUT

Subcorpus ?  
none (the whole corp...

Focus on ?  
rare  common

Minimum frequency ?  At least one alphanumeric ?  Only alphanumeric ?  
10

Keywords (single-words) settings

Reference corpus ?  
English Web 2015 (enTenTen15)

Reference subcorpus ?  
type to search

Maximum items ?  Attribute for keywords ?  
lemma

Terms (multi-words) settings

Reference corpus ?  
English Web 2015 (enTenTen15)

Reference subcorpus ?  
type to search

Maximum items ?

The keyword or term has to contain at least one letter or number to be included. Punctuation will not be included. Words such as 16-year-old or 3D will be included.



The keyword or term must consist of only letters and numbers. Words such as 16-year-old will not be included but 3D will be.


- Attribute for keywords ?
- word
  - lemma
  - tag
  - lempos
  - part of speech
  - lempos (lowercase)
  - lemma (lowercase)
  - word (lowercase)

# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Batch analysis by third-party extraction tools (Sketch Engine)

**KEYWORDS**   

** CHANGE VIEW OPTIONS**

- Show line numbers <sup>?</sup>
- Show counts <sup>?</sup>
- Show relative frequency <sup>?</sup>
- Show scores <sup>?</sup>
- Enable Wikipedia search <sup>?</sup>





# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



SINGLE-WORDS ✓ MULTI-WORDS ✓

↻ reference corpus: English Web 2015 (enTenTen15)

	Word	Frequency <sup>?</sup>		Relative freq. <sup>?</sup>		Score <sup>?</sup>		
		Focus	Reference	Focus	Reference			
1	supervisory	449	78,295	7,171.035	5.08	1,179.57	W	...
2	controller	504	430,053	8,049.447	27.904	278.52	W	...
3	derogation	24	6,467	383.307	0.42	270.71	W	...
4	pursuant	154	134,528	2,459.553	8.729	252.91	W	...
5	Article	342	379,370	5,462.124	24.616	213.27	W	...
6	OJ	19	6,869	303.451	0.446	210.59	W	...
7	pseudonymisation	12	174	191.653	0.011	190.5	W	...
8	processor	264	337,356	4,216.377	21.89	184.25	W	...
9	erasure	19	11,023	303.451	0.715	177.5	W	...
10	TFEU	14	4,429	223.596	0.287	174.46	W	...



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



## Batch analysis by third-party extraction tools (Sketch Engine)

7	pseudonymisation	12	174	191.653	0.011	190.5	w
8	processor	264	337,356	4,216.377	21.89	184.25	w
9	erasure	19	11,023	303.451	0.715	177.5	w
10	TFEU	14	4,429	223.596	0.287	174.46	w

<input checked="" type="radio"/>	<a href="#">Word Sketch (focus corpus)</a>	<a href="#">↗</a>
<input checked="" type="radio"/>	<a href="#">Word Sketch (reference corpus)</a>	<a href="#">↗</a>
<input type="checkbox"/>	<a href="#">Concordance (focus corpus)</a>	<a href="#">↗</a>
<input type="checkbox"/>	<a href="#">Concordance (reference corpus)</a>	<a href="#">↗</a>

## ARTICLES FROM WIKIPEDIA

- [1. Pseudonymization](#) [↗](#)
- [2. General Data Protection Regulation](#) [↗](#)
- [3. Data anonymization](#) [↗](#)
- [4. Data re-identification](#) [↗](#)



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



SINGLE-WORDS ✓

MULTI-WORDS ✓



reference corpus: English Web 2015 (enTenTen15)

	Word	Frequency <sup>?</sup>		Relative freq. <sup>?</sup>		Score <sup>?</sup>		
		Focus	Reference	Focus	Reference			
1	supervisory authority	328	2,888	5,238.529	0.187	4,414.09	W	...
2	third country	79	9,100	1,261.719	0.59	794.16	W	...
3	lead supervisory authority	47	1	750.643	0	751.64	W	...
4	personal data	51	2,546	814.527	0.165	700.02	W	...
5	natural person	45	3,704	718.701	0.24	580.4	W	...
6	international organisation	46	5,039	734.672	0.327	554.8	W	...
7	data subject	34	442	543.018	0.029	529.2	W	...
8	competent supervisory authority	33	76	527.047	0.005	525.94	W	...
9	personal data breach	30	86	479.134	0.006	477.74	W	...
10	protection officer	31	1,109	495.105	0.072	463.22	W	...

Rows per page:

20



1-20 of 46



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



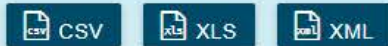
Batch analysis by third-party extraction tools (Sketch Engine)



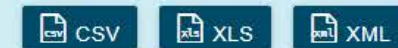
Download

↓ DOWNLOAD

Keywords



Terms



Or save the current view as



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Batch analysis by third-party extraction tools (Sketch Engine)

	A	B	C	D	E	F
1	Term	Score	Freq	Ref freq	Rel freq	Rel ref freq
2	supervisory	1179.57	449	78295	7171.035	5.08
3	controller	278.52	504	430053	8049.447	27.904
4	derogation	270.71	24	6467	383.307	0.42
5	pursuant	252.91	154	134528	2459.553	8.729
6	Article	213.27	342	379370	5462.124	24.616
7	OJ	210.59	19	6869	303.451	0.446
8	pseudonymisation	190.5	12	174	191.653	0.011
9	processor	184.25	264	337356	4216.377	21.89
10	erasure	177.5	19	11023	303.451	0.715



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Batch analysis by third-party extraction tools (Sketch Engine)

	A	B	C	D	E	F
1	Term	Score	Freq	Ref freq	Rel freq	Rel ref freq
2	supervisory authority	4414.09	328	2888	5238.529	0.187
3	third country	794.16	79	9100	1261.719	0.59
4	lead supervisory authority	751.64	47	1	750.643	0
5	personal data	700.02	51	2546	814.527	0.165
6	natural person	580.4	45	3704	718.701	0.24
7	international organisation	554.8	46	5039	734.672	0.327
8	data subject	529.2	34	442	543.018	0.029
9	competent supervisory authority	525.94	33	76	527.047	0.005
10	personal data breach	477.74	30	86	479.134	0.006



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



## Batch analysis by third-party extraction tools (Sketch Engine)

**CONCORDANCE**

cql [term(2,856642)] 328 (5,238.53 per million)

Details  Left context  KWIC  Right context

1	<input type="checkbox"/>	doc#0 the Union. </s><s> In cases involving both the controller and the processor, the competent lead <b>supervisory authority</b> should remain the supervisory au
2	<input type="checkbox"/>	doc#0 ith the controller and the processor, the competent lead supervisory authority should remain the <b>supervisory authority</b> of the Member State where the c
3	<input type="checkbox"/>	doc#0 ervisory authority of the Member State where the controller has its main establishment, but the <b>supervisory authority</b> of the processor should be consid
4	<input type="checkbox"/>	doc#0 main establishment, but the supervisory authority of the processor should be considered to be a <b>supervisory authority</b> concerned and that supervisory e
5	<input type="checkbox"/>	doc#0 uthority of the processor should be considered to be a supervisory authority concerned and that <b>supervisory authority</b> should participate in the coopera
6	<input type="checkbox"/>	doc#0 esentative should act on behalf of the controller or the processor and may be addressed by any <b>supervisory authority</b> . </s><s> The representative sho
7	<input type="checkbox"/>	doc#0 act or standard contractual clauses which are adopted either directly by the Commission or by a <b>supervisory authority</b> in accordance with the consisten
8	<input type="checkbox"/>	doc#0 ts responsibility. </s><s> Each controller and processor should be obliged to cooperate with the <b>supervisory authority</b> and make those records, on req
9	<input type="checkbox"/>	doc#0 measures in terms of available technology and costs of implementation, a consultation of the <b>supervisory authority</b> should take place prior to the pro
10	<input type="checkbox"/>	doc#0 personal data breach has occurred, the controller should notify the personal data breach to the <b>supervisory authority</b> without undue delay and, where



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



## Batch analysis by third-party extraction tools (Sketch Engine)

**Individual accounts**

Trial	ELEXIS-funded academic	Academic	Freelance (self-employed)
<b>Free</b> <i>for 30 days</i>	<b>0€</b> <i>from 2018 to 2022</i>	<i>from</i> <b>4.83€</b> <i>per month</i>	<i>from</i> <b>8.33€</b> <i>per month</i>
All functions	I'm a researcher, teacher or student from an <b>academic institution in the EU</b>	I'm a researcher, teacher or student	I'm a translator, terminologist, lexicographer, copywriter, marketing or branding specialist.
AND	AND	AND	AND
Featured corpora for all languages	I do not conduct any commercial activities.	I do not conduct any commercial activities or dictionary publishing.	I'm freelance (self-employed) and do conduct other kinds of commercial activity.
AND			
Corpus building for all languages	<a href="#">Gain access via your institution</a>	<a href="#">Calculate your price</a>	<a href="#">Calculate your price</a>
<a href="#">Start Trial</a>			





# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



## Batch analysis by third-party extraction tools (Prospector)

---

- developed by **Logrus Global Localization Cloud**
- uses a combination of proprietary linguistic algorithms and semantic relevancy measures to effectively identify terms, and advanced stemming technology to convert plurals and inflections to the base form
- in descending order of importance, on separate sheets of an Excel file: new terms, acronyms, and selected reference glossary
- uses the Corpus of Contemporary American English (COCA) as a reference corpus, which improves term ranking
- currently **free for freelancers**; subscription fees for corporate clients



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



## Batch analysis by third-party extraction tools (Prospector)

The tool supports the following formats:

- .txt – text files.
- .html, .htm – web pages.
- .docx, .doc – Microsoft Word documents.

GDPR\_CELEX\_3...79\_EN\_TXT.html

Select glossaries:

- Microsoft Generic Glossary
- Companies
- Cities
- Territories
- Acronyms

Beta version also supports xliff files and has more glossaries.



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Prospector (Beta)  
Source English  
terminology extractor  
with low noise and  
termbase integration



## Batch analysis by third-party extraction tools (Prospector)

	A	B	C
1	Count ▾	No ▾	Term
2	517	21	Member State
3	438	50	supervisory authority
4	96	18	European Parliament
5	96	34	Member State law
6	62	366	organisation
7	39	66	processing operation
8	37	283	data breach
9	33	301	data protection officer
10	32	20	processing activity

← → **Terms** Acronyms Microsoft Generic Glossary

794	1	307	written mandate
795	1	108	written statement
796	1	2	xml

← → **Terms** Acronyms Microsoft Generic Glossary



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Prospector (Beta)  
Source English  
terminology extractor  
with low noise and  
termbase integration



## Batch analysis by third-party extraction tools (Prospector)

	A	B	C	D
1	Cour	Nº	Acronym	
2	52	8	EC	
3	19	836	OJ	
4	18	5	EU	
5	13	17	TFEU	
6	4	163	EEC	
7	3	812	EUR	
8	2	1	EN	
9	2	573	TEU	
10	1	189	CSIRT	
11	1	700	EN-ISO	
12	1	835	HENNIS-PLASSCHAERT	
13	1	841	JHA	
14	1	507	ne	
15	1	834	SCHULZ	

Terms **Acronyms** Microsoft Generic Glossary



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Prospector (Beta)  
Source English  
terminology extractor  
with low noise and  
termbase integration



## Batch analysis by third-party extraction tools (Prospector)

	A	B	C
1	Cour ▾	No ▾	Microsoft Generic Glossary ▾
2	395	43	data subject
3	31	35	data protection
4	18	4	European Union
5	7	7	April
6	7	113	data processing
7	6	124	health status
8	6	381	Personal Data
9	5	237	electronic form
10	5	224	health care

Terms Acronyms **Microsoft Generic Glossary**

53	1	275	statistical procedure
54	1	144	user profile

Terms Acronyms **Microsoft Generic Glossary**



# EXTRACTING TERM CANDIDATES FOR THE GLOSSARY



Prospector (Beta)  
Source English  
terminology extractor  
with low noise and  
termbase integration



Batch analysis by third-party extraction tools (Prospector)



Termlode (Beta)



Memose (Beta)



Rigora



Prospector (Beta)



Termlode Trados  
Connector



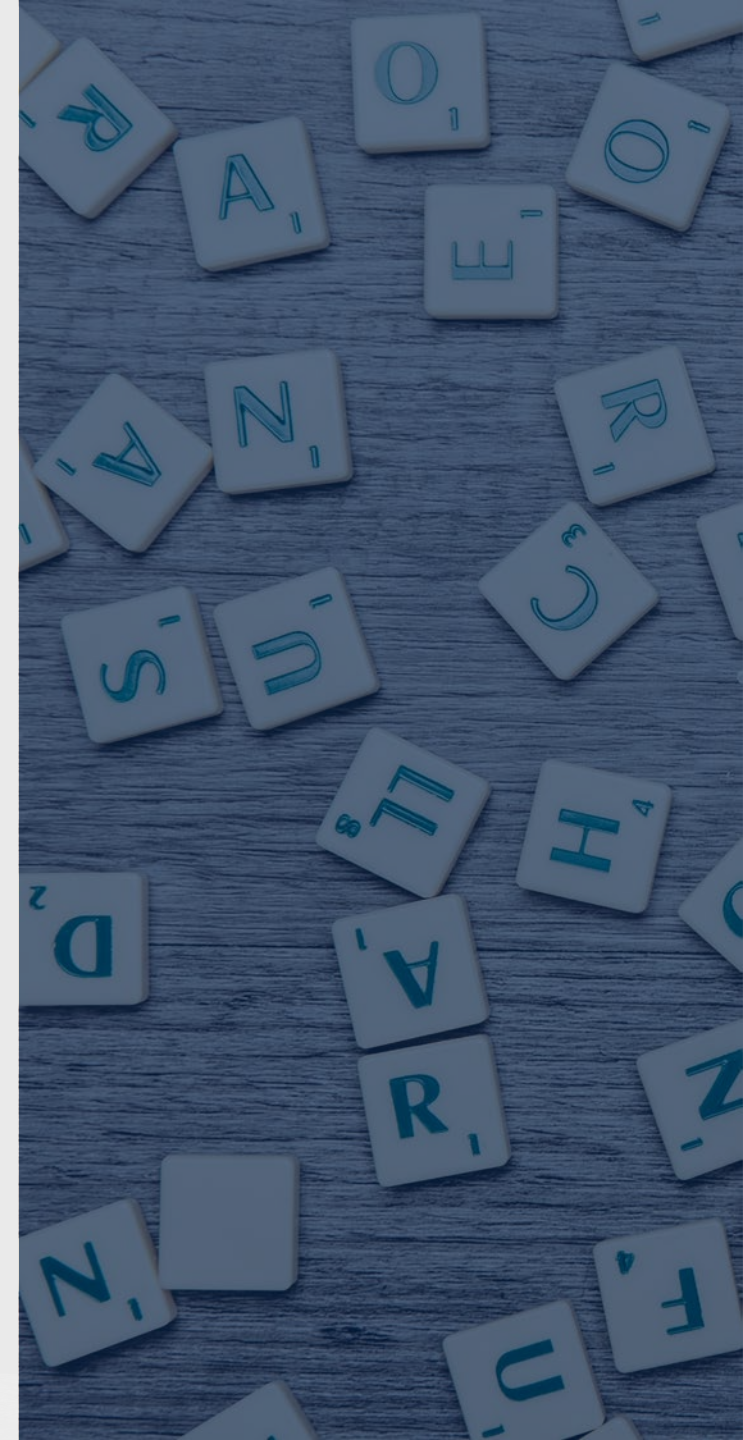
# CONSIDERATIONS FOR THE GLOSSARY MATRIX



Essential and optional attributes and metadata



Capturing and supplementing metadata



# CONSIDERATIONS FOR THE GLOSSARY MATRIX



Essential and optional attributes and metadata

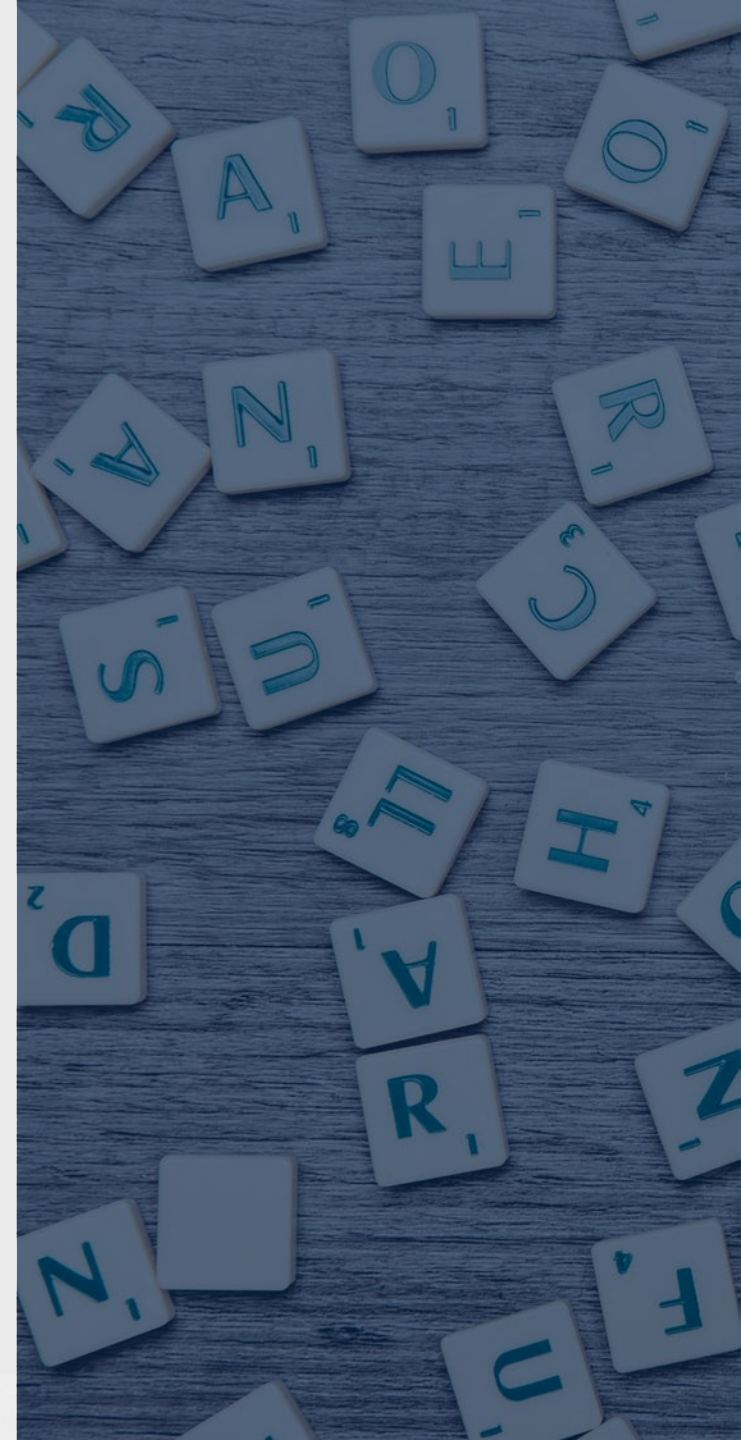
**E**xtract terms

**L**abel with metadata

**A**nalyze and review for consistency, suitability, and scope

**T**ermbase import and maintenance

**D**istribute to source and target teams





# CONSIDERATIONS FOR THE GLOSSARY MATRIX

	A	B	C	D	E	F	G	H	I	J	K	L
1	<b>Domain/ Subject Field</b>	<b>Subfield</b>	<b>Source Document</b>	<b>Context</b>	<b>Definition</b>	<b>Part of Speech</b>	<b>Originator Name</b>	<b>Creation Date</b>	<b>Updater Name</b>	<b>Update Date</b>	<b>Source Language Code, e.g. en-US</b>	<b>Optional metadata</b>
2												
3												<b>phraseologisms</b>
4												<b>usage labels</b>
5												<b>approval status: preferred deprecated, etc.</b>
6												<b>term cross-references: superordinate, subordinate, coordinate, synonyms, abbreviations, expanded forms, acronyms</b>



# CONSIDERATIONS FOR THE GLOSSARY MATRIX

M	N	O	P	Q	R	S	T	U	V
<b>Target Language Code</b>	<b>Gender</b>	<b>Number</b>	<b>Originator Name</b>	<b>Creation Date</b>	<b>Updater Name</b>	<b>Update Date</b>	<b>Approval Status</b>	<b>Language Comments</b>	<b>Optional metadata</b>
									<b>phraseologisms</b>
									<b>usage labels</b>
									<b>approval status:</b> <b>preferred</b> <b>deprecated, etc.</b>
									<b>term cross-references:</b> <b>superordinate,</b> <b>subordinate,</b> <b>coordinate,</b> <b>synonyms,</b> <b>abbreviations,</b> <b>expanded forms,</b> <b>acronyms</b>

# CONSIDERATIONS FOR THE GLOSSARY MATRIX



## Capturing and supplementing metadata

- **Context** from extraction spreadsheet or source document
- **Definitions** from source document or subject matter references

### Definitions

For the purpose of this Directive, the following **definitions** shall apply:

- (1) ‘consumer’ means any natural person who, in contracts covered by this Directive, is acting for purposes which are outside his trade, business, craft or profession;
- (2) ‘trader’ means any natural person or any legal person, irrespective of whether privately or publicly owned, who is acting, including through any other person acting in his name or on his behalf, for purposes relating to his trade, business, craft or profession in relation to contracts covered by this Directive;

# ANALYZING THE TERM CANDIDATE SELECTION

**E**xtract terms

**L**abel with metadata

**A**nalyze and review for consistency, suitability, and scope

**T**ermbase import and maintenance

**D**istribute to source and target teams



# ANALYZING THE TERM CANDIDATE SELECTION



## Quantitative and qualitative criteria

### Quantitative:

- review frequency
- review score

### Qualitative:

- review delimiting semantic features based on the concept
- distinguish key words, subsets, full term  
(e.g. *author, authority, supervisory authority*)



# ANALYZING THE TERM CANDIDATE SELECTION



## Determining the scope

- ❑ Determine a cut-off score or frequency
- ❑ Consider workload and budget
- ❑ Balance impact on QA processes:
  - more terms: consistency improvement, but higher workload
  - fewer terms: lower workload, possibly at the expense of quality



# ANALYZING THE TERM CANDIDATE SELECTION



## Establishing term relationships

- ✓ Superordinate
- ✓ Coordinate
- ✓ Subordinate

Coordinate terms: terms sharing identical subsets

213	2	724	data protection audit
214	2	359	data protection law
215	2	731	data protection legislation
216	2	595	data protection policy
217	2	280	data protection principle



# ANALYZING THE TERM CANDIDATE SELECTION



## Establishing term relationships

- ✓ Superordinate
- ✓ Coordinate
- ✓ Subordinate

**Superordinate vs. subordinate:**  
compare keywords with multi-word expressions

	A	B
1	<b>Term</b>	<b>Term</b>
2	supervisory	supervisory
3	controller	controller
4	derogation	derogation
5	pursuant	pursuant
6	Article	Article
7	OJ	OJ
8	pseudonymisation	pseudonymisation
9	processor	processor
10	erasure	erasure
11	TFEU	TFEU
12	processing	processing
13	paragraph	paragraph
14	Directive	Directive
15	infringement	infringement
16	lawfulness	lawfulness
17	subparagraph	subparagraph
18	rectification	rectification
19	Official	Official
20	enforceable	enforceable

	A	B
1	<b>Term</b>	<b>target</b>
2	supervisory authority	target placeholder
3	third country	target placeholder
4	lead supervisory authority	target placeholder
5	personal data	target placeholder
6	natural person	target placeholder
7	international organisation	target placeholder
8	data subject	target placeholder
9	competent supervisory authority	target placeholder
10	personal data breach	target placeholder
11	protection officer	target placeholder
12	data breach	target placeholder
13	protection impact	target placeholder
14	data protection	target placeholder
15	protection impact assessment	target placeholder
16	consistency mechanism	target placeholder
17	such processing	target placeholder
18	draft decision	target placeholder
19	main establishment	target placeholder
20	judicial remedy	target placeholder

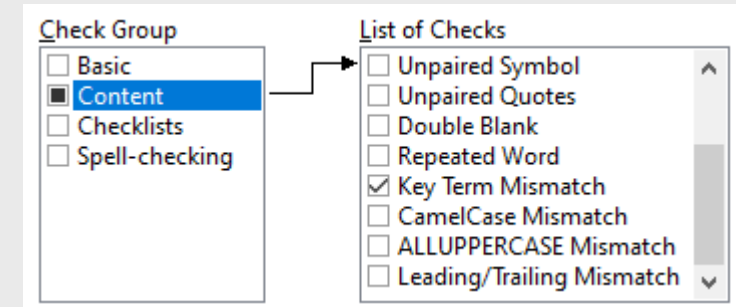




# ANALYZING THE TERM CANDIDATE SELECTION



**Superordinate vs. subordinate:**  
compare keywords with multi-  
word expressions



## Establishing term relationships

- ✓ Superordinate
- ✓ Coordinate
- ✓ Subordinate

Project Properties

Files Settings

Glossary List

	Priority	C	Type	File/Folder Name
★	Medium		Tab-delimited Text	xbench_keywords_gdpr_en.txt
•	Low		Tab-delimited Text	xbench_terms_gdpr_en.txt

Add... Properties... Remove

Priority: + -

Key Terms

Ongoing Translation



# ANALYZING THE TERM CANDIDATE SELECTION



## Establishing term relationships

- ✓ Superordinate
- ✓ Coordinate
- ✓ Subordinate

99	Key Term Mismatch (authority / authority)	
100	<i>xbench_terms_gdpr_en.txt (23)</i>	public <b>authority</b>
101	<i>xbench_terms_gdpr_en.txt (20)</i>	official <b>authority</b>
102	<i>xbench_terms_gdpr_en.txt (1)</i>	supervisory <b>authority</b>
103	<i>xbench_terms_gdpr_en.txt (3)</i>	lead supervisory <b>authority</b>
104	<i>xbench_terms_gdpr_en.txt (26)</i>	exercise of official <b>authority</b>
105	<i>xbench_terms_gdpr_en.txt (8)</i>	competent supervisory <b>authority</b>

46	Key Term Mismatch (data / data)	
47	<i>xbench_terms_gdpr_en.txt (11)</i>	<b>data</b> breach
48	<i>xbench_terms_gdpr_en.txt (7)</i>	<b>data</b> subject
49	<i>xbench_terms_gdpr_en.txt (4)</i>	personal <b>data</b>
50	<i>xbench_terms_gdpr_en.txt (13)</i>	<b>data</b> protection
51	<i>xbench_terms_gdpr_en.txt (9)</i>	personal <b>data</b> breach

180	Key Term Mismatch (public / public)	
181	<i>xbench_terms_gdpr_en.txt (46)</i>	<b>public</b> health
182	<i>xbench_terms_gdpr_en.txt (32)</i>	<b>public</b> interest
183	<i>xbench_terms_gdpr_en.txt (38)</i>	<b>public</b> security
184	<i>xbench_terms_gdpr_en.txt (23)</i>	<b>public</b> authority

Establishing source term relationship impacts target QA !





# Termbase Integration

# TERMBASE INTEGRATION

**E**xtract terms

**L**abel with metadata

**A**nalyze and review for consistency, suitability, and scope

**Te**rmbase import and maintenance

**D**istribute to source and target teams



# TERMBASE INTEGRATION

SDL MultiTerm Convert - Specify Column Header (5/9)

This screen allows you to specify the type of each column header field contained in the input file. You need to specify the language for language fields.

Available column header fields:

- Domain/Subject Field
- Subfield
- Source Document
- Context
- Definition
- Part of Speech
- Originator Name
- Creation Date
- Updater Name
- Update Date
- Source Language Code, e.g. en-US
- Optional metadata
- Target Language Code
- Gender
- Number
- Approval Status
- Language Comments

Language field  
English (United States)

Descriptive field  
Text

Concept ID

Creation Date

Updater Name

Update Date

Source Language Code, e.g. en-US

Optional metadata

Target Language Code

Gender

Number

Approval Status

Language Comments

Descriptive field  
Picklist

Concept ID

Available column header fields:

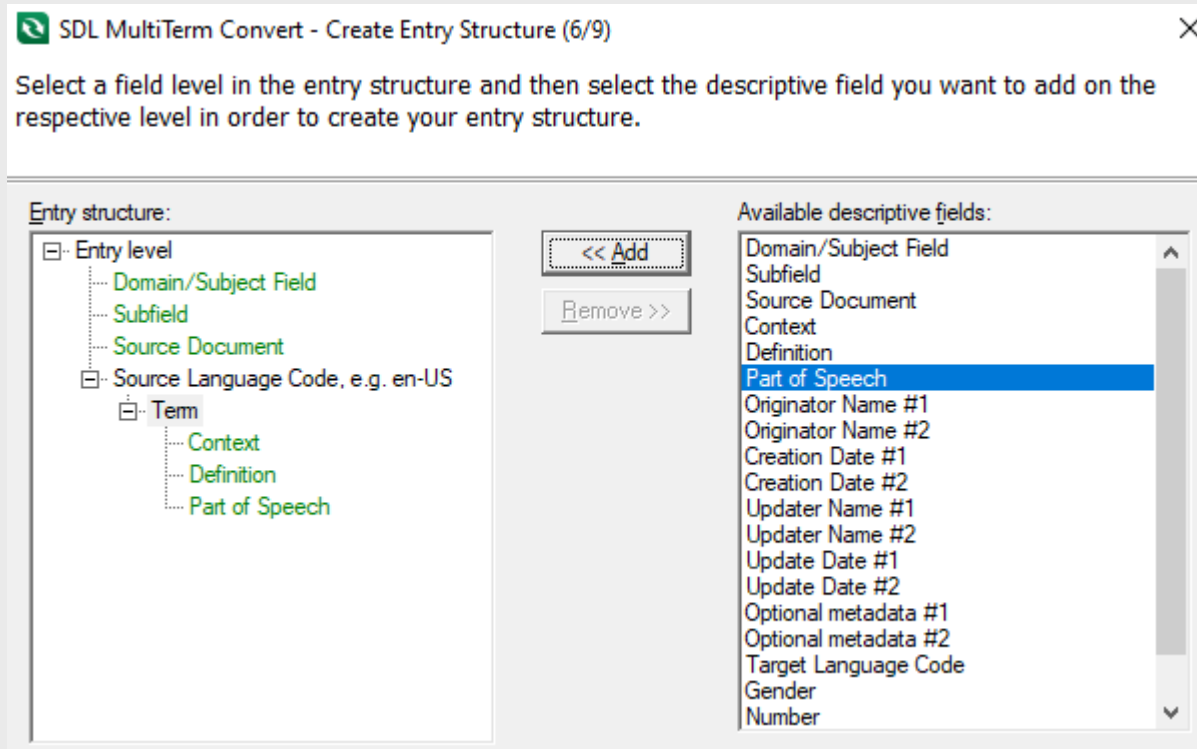
- Domain/Subject Field
- Subfield
- Source Document
- Context
- Definition
- Part of Speech
- Originator Name
- Creation Date
- Updater Name
- Update Date
- Source Language Code, e.g. en-US
- Optional metadata
- Target Language Code

Language field  
English (United States)

Descriptive field  
Text



# TERMBASE INTEGRATION



## Termbase Entry Field Levels

There are three levels at which you can create fields in a termbase entry: Entry, Language and Term.

- **Entry level** – This is the top level and holds information that applies to the whole termbase entry.
  - **Language level** – Language level fields hold information that is relevant to a language.
    - **Term level** – Term level fields hold data that is specific to the term.

Conversion alternative\*:  
(in SDL Appstore)



Q&A



# REFERENCES

- Pavel Terminology Tutorial  
*<http://linguistech.ca/pavel/>*
- Patricia Brenes: Pavel's Key Points  
*<http://inmyownterms.com/pavels-key-points/>*
- TermNet  
*<https://www.termnet.org/index.php>*
- B. I. Karsch's Blog  
*<http://bikterminology.com/blog/>*
- Jost Zetzsche's Tool Box Journal  
*<https://www.internationalwriters.com/toolkit/index.html>*
- XTM  
*<https://xtm.cloud/knowledge-base/how-to-extract-terminology/>*
- Sketch Engine  
*<https://www.sketchengine.eu/>*
- Prospector  
*<https://cloud.logrusglobal.com/#prospector>*
- Glossary Converter  
*<https://appstore.sdl.com/language/app/glossary-converter/195/>*
- Sameh Ragab - AST-04:  
A New Advanced Approach to Terminology Management and Compilation  
*<https://event.ata61virtual.com/agenda>*





# LINKS

- Complete set of presentation slides (available on Oct. 23, 2020):  
<https://www.argosmultilingual.com/new-client-new-terms-glossary-101>
- Coffee Break interview:  
<https://www.argosmultilingual.com/coffee-break-talking-translations-german-language-leading>
- Blog segment: Life as a QA Specialist  
<https://www.argosmultilingual.com/blog/life-as-a-qa-specialist>



# THANK YOU

---



[ARGOSMULTILINGUAL.COM](https://www.argosmultilingual.com)