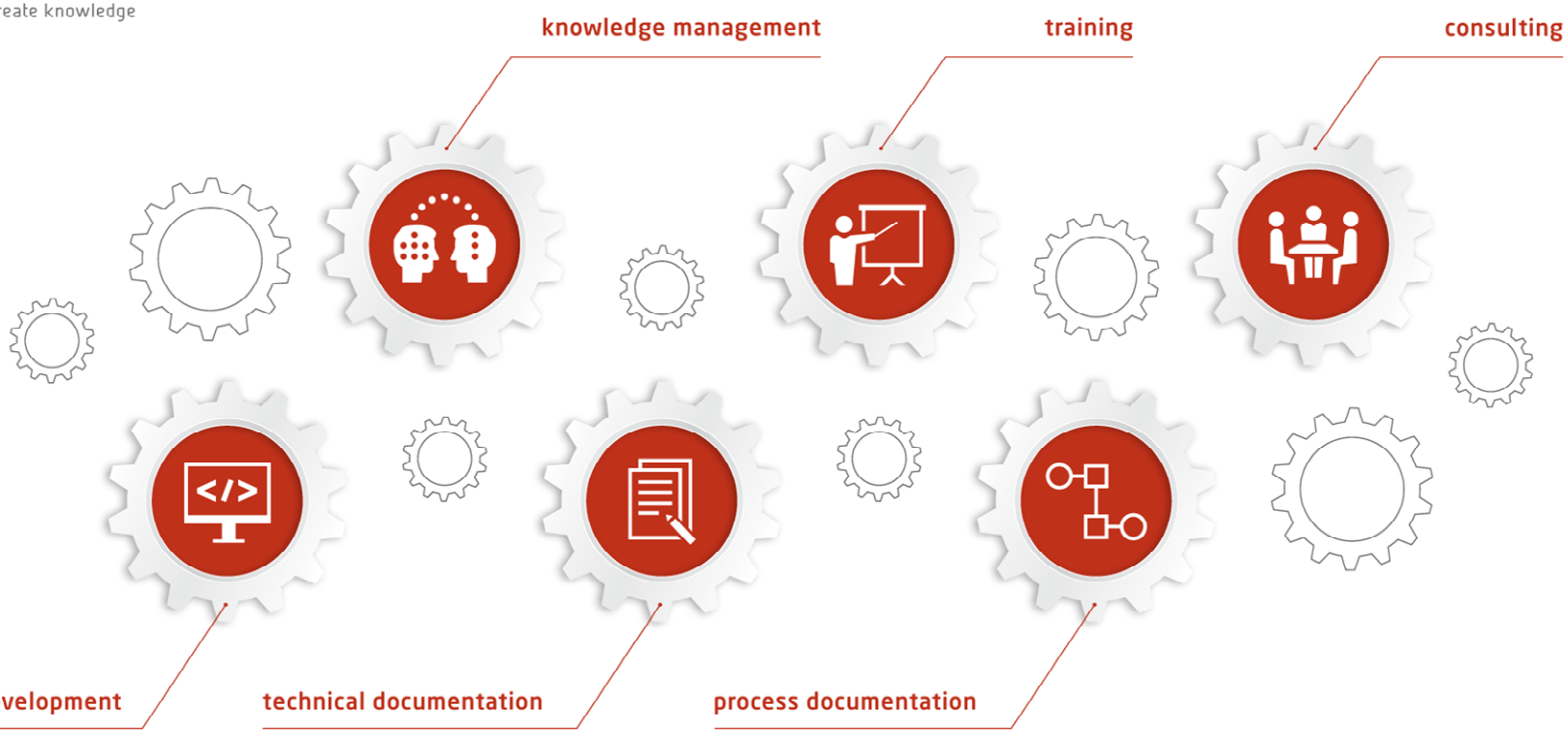
parson

Charming the word snake – Terminology work and language checking with Python

Maximilian Rosin, Esther Strauch,
parson AG

parson
we create knowledge



www.parson-europe.com

parson

Your friendly snake charmers

Esther Strauch
Technical Communicator

Maximilian Rosin
Technical Consultant

parson

Part 1 - Introduction

Part 2 - Terminology extraction

Part 3 - Checking writing rules

Part 4 - Outlook and discussion

parson

Scope

- Implement powerful tools in just a few days, using only free open-source software.
- Find a starting point for terminology work.
- Start with a specific research question.

parson

Survey questions

1. How do you work with terminology?
   a) I have not started yet.
   b) I work with Excel lists.
   c) I use a terminology software.
2. Do you have any experience with programming Python?
   a) Yes.
   b) No, but I am familiar with other languages.
   c) No, I have no programming experience at all.

parson

Part 1 - Introduction

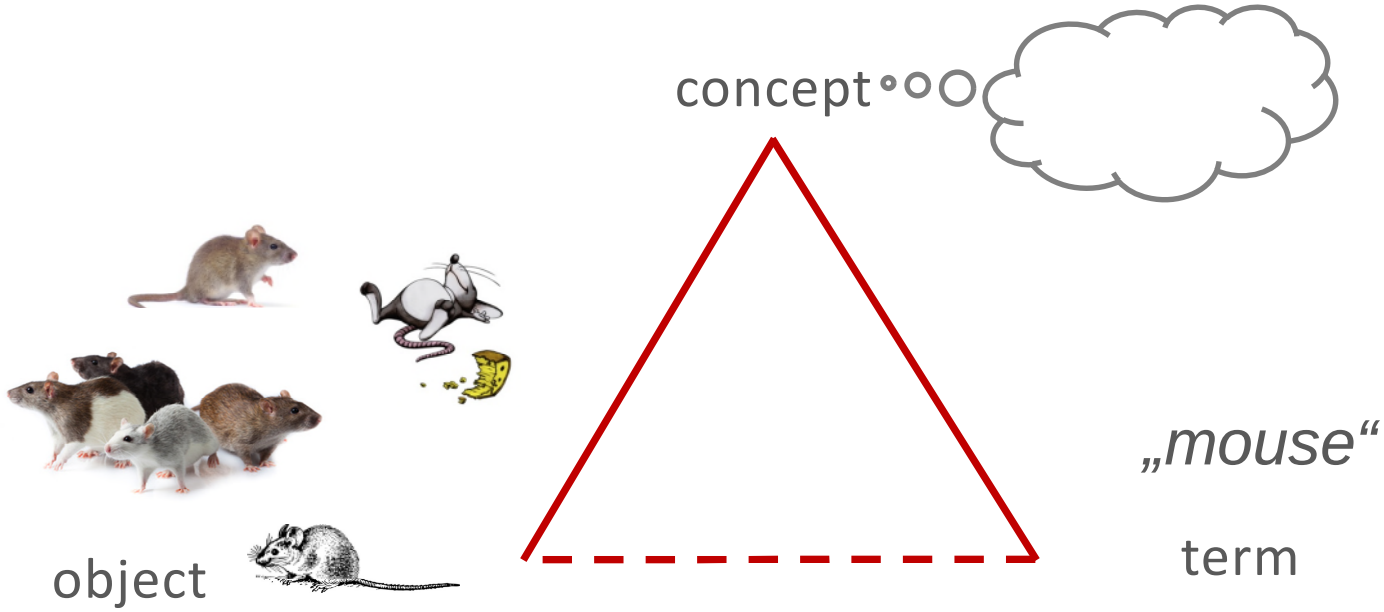
Part 2 - Terminology extraction

Part 3 - Checking writing rules

Part 4 - Outlook and discussion

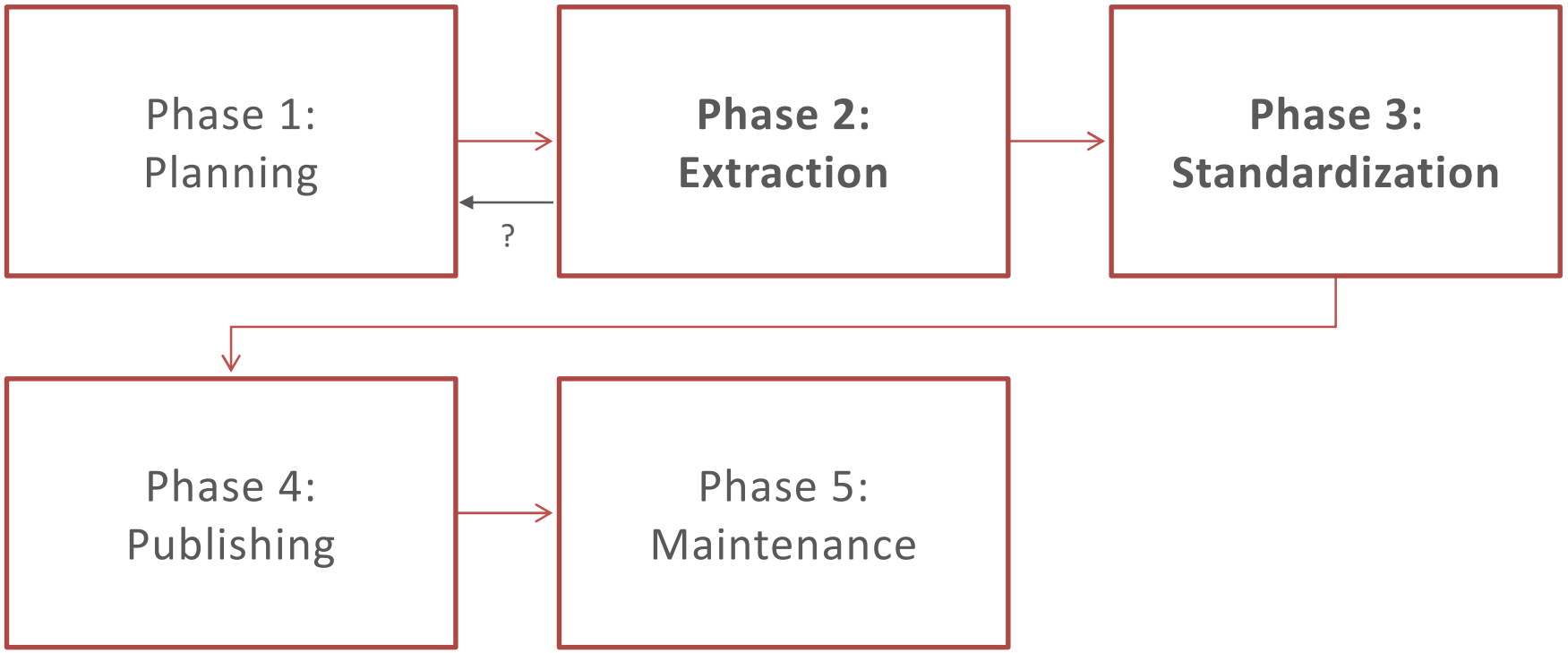
parson

Terminology: Triangle of Meaning



parson

Terminology projects



parson

Metadata useful for Standardization

Concept level
ID
Subject area
Definition
Source

Term level
Language
Term
Status
Source
Part of speech
Context
Comment

Concept level
ID
Subject area
Definition
Source

Term level
Language
Term
Status
Source
Part of speech
Context
Comment

Standardization

Generate automatically

- Term, source, context, POS, definition

Complete with terminology work

- Concept, subject field, status, relations between concepts

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Concept Level						Language Level			Term Level					
2	Concept ID	Subject field	Project	Status	ID Related concept	Relation	Definition	Comments	Source	Term	Usage	Source	Comment	Context	POS
3	1	A	Terminology	In review	2	parent	Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.			XX	deprecated	http://www.someurl.de			noun
4	1	A	Terminology	In review	2	parent	Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.			xx	preferred	http://www.someurl.de			noun
5	2	A	Terminology	In review	1	child	At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.			YY	preferred	http://www.someurl.de			noun
	2	A	Terminology	In review	1	child	At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus			yy	deprecated	http://www.someurl.de			noun

parson

Live programming

⇨ Jupyter notebook

parson

Summary: Terminology extraction

Feed text into spacy‘s language model.

Use language model to extract one-word terms and part-of-speech tags.

Add sample sentences for context.

Add definitions from 3rd-party source where possible.

Write results into CSV file for further processing.

parson

Questions?

parson

Part 1 - Introduction

Part 2 - Terminology extraction

Part 3 - Checking writing rules

Part 4 - Outlook and discussion

parson

Live programming

⇨ Jupyter notebook

parson

Summary: Checking writing rules

Feed text into spacy‘s German language model.

Find forbidden words based only on their canonical form.

Find overly long words by counting syllables and letters.

Find complicated sentences by counting words and commas.

Find nominalizations by examining POS tags and suffixes.

parson

Part 1 - Introduction

Part 2 - Terminology extraction

Part 3 - Checking writing rules

Part 4 - Outlook and discussion

parson

Advanced use cases

- Building domain-specific terminology from web content
- Integration scenarios with Docs-as-Code or CCMS
- Automated reports on text quality

parson

Limitations of DIY approach

- No real-time-checks
- Not integrated in editor
- No propositions of allowed terms
- No „linguistic intelligence“

parson

Questions?

parson

References

- Ann-Cathrin Mackenthun: *Terminology management on a budget* (https://www.parson-europe.com/en/knowledge-base/terminology-management-on-a-budget)
- Al Sweigart: *Automate the Boring Stuff with Python* (https://automatetheboringstuff.com/)
- Python 3: The programming language (https://www.python.org/)
- Anaconda: Package manager for data science (https://www.anaconda.com/)
- Jupyter Notebook: Creating documents with live code (https://jupyter.org/)
- spacy: NLP library (https://www.spacy.io)
- tabulate: Formatting output as tables (https://github.com/astanin/python-tabulate)
- json: Reading and writing JSON files (https://docs.python.org/3/library/json.html)
- csv: Reading and writing CSV files (https://docs.python.org/3/library/csv.html)

parson

How did you like our tutorial today?

Maximilian Rosin

maximilian.rosin@parson-europe.com

Esther Strauch

esther.strauch@parson-europe.com

parson AG

www.parson-europe.com

we create knowledge

parson

Maximilian Rosin
+49 (0)151 11 96 38 18
maximilian.rosin@parson-europe.com

Esther Strauch
+49 (0)151 19 32 84 15
esther.strauch@parson-europe.com

parson AG
Reinbeker Redder 94
21031 Hamburg
+49 (0)40 7200 500-0
contact@parson-europe.com
www.parson-europe.com